Experimental Education Series

Edited by M. V. O'SHEA

THE FUNDAMENTALS OF STATISTICS

EXPERIMENTAL EDUCATION SERIES

EDITED BY M. V. O'SHEA

HOW TO EXPERIMENT IN EDUCATION.

By WILLIAM A. MCCALL, Associate Professor of Education, Teachers College, Columbia University.

SPECIAL TALENTS AND DEFECTS.

By LETA S. HOLLINGWORTH, Associate Professor of Education, Teachers College, Columbia University.

FITTING THE SCHOOL TO THE CHILD.

By ELISABETH A. IRWIN, Psychologist, Public Education Association of New York City, and LOUIS A. MARKS, Member Board of Examiners, Board of Education, New York City.

THE FUNDAMENTALS OF STATISTICS.

By L. L. THURSTONE, Associate Professor of Psychology, University of Chicago.

THE FUNDAMENTALS OF STATISTICS

BY

L. L. THURSTONE, M.E., PH.D.
Associate Professor of Psychology,
University of Chicago

New York
THE MACMILLAN COMPANY

Set up and electrotyped. Published February, 1925. Reprinted December, 1925; May, 1927; June, 1928; July, 1930; May, 1931; August, 1935; April, 1937; September, 1938; June, 1941; May, December, 1944; September, 1945; March, 1946; October, 1948; June, 1949.

PRINTED IN THE UNITED STATES OF AMERICA

Editor's Introduction

IT is impossible for any person to read very much of present-day educational literature with pleasure and understanding unless he is acquainted to some extent with the method and the terminology employed in conducting and presenting the results of statistical investigation. The character of educational literature has undergone a fundamental change during the past ten or fifteen years. Doubtless some of the readers of these lines can remember the time when articles and books relating to educational values, methods, or administration rarely, if ever, contained tables of data or graphs showing the distribution of the facts bearing upon any phase of educational procedure. In those days a teacher hardly ever — probably never — came across in her professional reading such terms as *correlation coefficient*, *median*, *mode*, *variability*, *frequency tables*, *frequency surface*, *standard deviation*, *probability curve*, *percentile ranks*, and so on. But choose at random twenty-five books or articles on educational subjects that have recently appeared and that are regarded as up-to-date, and it will undoubtedly be found that the content of twenty of them is based upon investigations involving statistical method, and that the conclusions are phrased in terms which can be compre-

hended only when one is familiar with statistical modes of organizing and presenting data.

The time is passing rapidly when one can safely discuss most educational subjects without having first made a statistical investigation of the topics which he treats. Formerly educational writers discussed problems of values, methods, and administration without any attempt to be statistically sound and consistent in their premises or conclusions. They relied upon "reason" to guide them accurately in matters pertaining to educational procedure. They analyzed problems under consideration, and on the basis of experience, logic, or "common sense" they deduced principles without deeming it necessary to determine to what extent these principles would hold good in the concrete situations to which they related. But to-day we have slight confidence in the validity or value of principles arrived at in this fashion. We demand of anyone who presumes to speak authoritatively upon educational themes that he shall first have gathered the data pertaining to the themes in an accurate way, and that he shall then have treated the data so as to clarify obscure matters, eliminate error from his conclusions, and show how his facts are distributed.

It is universally recognized that all the data relating to educational problems are complex; the situations out of which they have sprung are complicated, and in general any factor pertaining to any aspect of education is intimately bound up functionally with other factors, so that it is not possible to de-

scribe the characteristics or measure the force of the factor in question unless it can be isolated for purposes of measurement from the factors with which it is ordinarily associated. The untrained student cannot accomplish this in treating the problems which he wishes to investigate. Even if he can secure data by proper modes of procedure, he does not know how to organize the data and present conclusions so as to reveal their trend and meaning. For this reason it is essential that, before attempting investigation, he should become familiar with methods of procedure that will enable him to avoid error in the treatment of his data and to derive principles which the data should yield.

Unfortunately the books that deal with statistical method in the treatment of educational data have seemed to the novitiate in educational investigation to be technical and forbidding. The very appearance of most books on statistics has been enough to deter the novice from attempting to master this subject. The present writer believes that Mr. Thurstone's volume will be more acceptable than the typical textbook on statistical method to those who wish to become familiar with the method so that they can read educational literature understandingly and agreeably, and so that they may participate according to their resources in the investigation of educational problems. The author has succeeded in discussing every detail of statistical procedure in such a simple and clear way and in such terminology that it can be comprehended by a teacher or investigator who has not

been able to devote much time and energy to statistical study. Every essential principle is illustrated by concrete, impressive instances. The book might be regarded as a primer or introduction to the fundamental principles of statistics. The author has had extensive experience in teaching statistical method to students in educational psychology. He has made himself familiar with the difficulties which novices encounter in understanding the logical implications of statistics, and he has acquired exceptional skill in making these implications clear and intelligible.

The book may be heartily commended for its simplicity and clarity to all who wish to contribute to educational investigation as well as to those who wish to understand the results of investigations made by others.

M. V. O'SHEA.

THE UNIVERSITY OF WISCONSIN.

Preface

THIS textbook in statistics is the result of seven years of teaching the fundamental principles of statistics and mental measurements to classes of about thirty graduate students annually. I have found that the majority of the students who enter psychology in their graduate work come into it from undergraduate majors in economics, literature, languages, psychology, and other unmathematical subjects. They have dodged the college course in mathematics, and they seldom have any occasion to keep fresh their high school mathematics. For this reason it has been necessary to assume on their part very little knowledge of even the fundamentals of high school algebra, but this lack is often compensated for by a keen critical ability in the logic of the subject. This critical attitude toward statistical work I have always tried to encourage, and I have discouraged, as far as possible, the blind substitution of numbers into formulæ. It happens often that a student learns readily to calculate correlation coefficients and to draw pretty charts, but unless he understands the logical implications of his statistical work, more harm than good has been done. It has been my aim throughout these lessons to make clear the meaning and implications of statistical procedures.

I owe to my students much of the form of the explanations, since I have included in these lessons those particular examples, methods, and teaching tricks which have seemed to be most successful in explaining the subject to unmathematically inclined students.

I have borrowed extensively from various textbooks on statistics. The correlation data sheet is a modification of Thorndike's correlation table. I wish to acknowledge the use of *Figures 30* and *38* and some of the related text on percentiles and on the correlation table from my articles in the *Journal of Educational Research*, for which the publishers have kindly given their consent. I have adapted *Tables 19* and *20* from the extensive tables prepared by W. F. Sheppard and published in "Tables for Statisticians and Biometricians," edited by Karl Pearson. The two tables are brief, but they should serve the purposes of the beginner in statistical work. He will refer to Sheppard's tables for more accurate determinations.

I hope that this manual may prove useful not only to the beginning student in statistics but also to those workers in the field of mental measurement who do not feel at home in the logical interpretation of their statistical work. Many of the outlines and sample problems have been so arranged that they should prove helpful also for reference purposes. After the student has become familiar with the logical aspects of the measures of central tendency and variability and the meaning of the correlation table, he should be able to pursue his statistical studies further in the

more comprehensive textbooks, such as those of Yule, Bowley, Elderton, Brown and Thompson, and Kelley. The one outstanding textbook in statistics for the student whose training in mathematics is limited is that of Yule. It is so exceedingly well done that it would be folly for me to attempt anything more than an introduction to it.

Teachers and students who may use this manual will confer a favor, for which I shall be grateful, if they will call my attention to errors either in the text or in the arithmetical work.

L. L. THURSTONE.

CHICAGO, July, 1924.

Contents

List of Tables

List of Figures

THE FUNDAMENTALS OF STATISTICS

The Fundamentals of Statistics

Chapter One

The Frequency Table

When we have a collection of facts in numerical form, the first statistical task is usually to classify the data in some way. Suppose that a mental test has been given to three hundred students and that the papers have been scored. Some one inquires for the score of Mr. Jones, and we find that his score is 79. Is that a high score or a low score? We cannot answer that unless we know how many of the students scored *above* 79 and how many of them scored *below* 79. If all the other students scored below 79, then Jones' score is high; but if all the other students scored above 79, then Jones' score is low. If 150 of the 300 students scored above 79 and 150 of them below 79, then Jones has an average or ordinary score. This will suffice to show that it is not enough to give the test and score the papers; we must also classify the scores so that we may know how many students scored in the nineties, how many in the eighties, and so on. Such a table is called a *frequency table*.

An intelligence test was given to the freshmen at Swarthmore College. In *Table 1* we have a list of scores. Each number is the score of a student. If we want to know how many students obtained scores above 79, it is necessary to look through the whole

62	129	95	123	81	93	105	95	96	80
123	60	72	86	108	120	57	113	65	108
109	84	121	60	84	128	100	72	119	103
77	91	51	100	63	107	76	82	110	63
104	107	63	117	116	86	115	62	122	92
69	116	82	95	72	121	52	80	100	85
94	84	123	42	90	91	81	116	73	79
100	79	101	98	110	95	67	77	91	95
79	92	73	83	74	125	101	82	71	75
125	56	86	98	106	72	117	89	99	86
87	90	80	131	102	117	98	74	101	82
110	137	99	65	113	85	82	90	102	57
139	74	149	114	74	102	69	134	78	106
75	106	85	103	78	106	102	94	108	90

Table 1. Scores of Swarthmore College freshmen in an intelligence test

table. That is not necessary when the data are arranged in the form of a frequency table, as shown in *Figure 1*.

The frequency table is prepared as follows:

1. Arrange a *data sheet*[1] with the three headings *Score*, *Tabulation*, and *Frequency*, as shown in *Figure 1*.

2. Read off the scores in *Table 1* and for each one record a check mark as shown in *Figure 1*. The subsequent counting is facilitated if every fifth mark is made slanting across the preceding four checks, as shown in *Figure 1*.

[1] A data sheet is a sheet ruled with vertical columns for recording numerical or other data. In recording facts on a data sheet, be sure to label each column.

3. Add the check marks in each *row* and record the sums under *Frequency*.

Score	Tabulation	Frequency
0 – 9		
10-19		
20-29		
30-39		
40-49	𝍸	1
50-59	𝍸	5
60-69	𝍸 𝍸 II	12
70-79	𝍸 𝍸 𝍸 𝍸 I	21
80-89	𝍸 𝍸 𝍸 𝍸 III	23
90-99	𝍸 𝍸 𝍸 𝍸 III	23
100-109	𝍸 𝍸 𝍸 𝍸 𝍸	25
110-119	𝍸 𝍸 IIII	14
120-129	𝍸 𝍸 I	11
130-139	IIII	4
140-149	I	1
150-159		
160-169		
Total number of students =		140

Figure 1. Frequency table of scores in intelligence test for Swarthmore freshmen

4. Add the frequency *column*. This sum is the total number of cases and should agree with the number of scores in *Table 1*.

It is now possible to answer such questions as these:

1. How many students obtained scores between 70 and 79? (21)
2. How many students obtained scores between 60 and 69? (12)
3. How many students obtained scores above 99? (55)
4. How many students obtained scores below 70? (18)
5. How many students obtained scores between 20 and 49? (1)
6. What per cent of the freshman class obtained scores between 80 and 89? (16%)
7. What per cent of the freshman class obtained scores between 40 and 79? (28%)
8. Is a score of 95 high, average, or low? (Average)
9. What per cent of the class exceeded the score of 89? (56%)

All these questions and others of the same kind can be answered by referring to the frequency table.

A *variable* is any quantity which can have different numerical values. It is any varying quantity. Examples of variables are ages, birth- and death-rates, prices, wages, barometer readings, rainfall records, and city populations. The scores obtained in intelligence tests constitute a variable. We may consider as a variable the scores made by the different persons in a group. We may also consider as a variable the scores that a single person makes on the same test on different occasions.

Variables may be classified as *continuous* and *discontinuous*. If it is possible for a variable to change its numerical value by infinitesimally small degrees, it is called a continuous variable. If this is not possible, the variable is called discontinuous. Temperature, for instance, changes from 68 degrees to a new value, such as 69 degrees, by passing through all the intermediate values. Therefore temperature is considered a continuous variable. The number of freight cars in a train is a discontinuous variable because this variable does not change by passing through all intermediate values. If the train is 68 cars long, it cannot be made 69 cars long by passing through the intermediate values of, let us say, $68\frac{1}{3}$ cars and $68\frac{5}{7}$ cars.

The *range* is the difference between the maximum and minimum values of the variable in any series. In this group of Swarthmore freshmen the range of intelligence test scores is 107 because this is the difference between the highest and lowest scores in the freshman class (149 and 42 respectively).

The *class interval* is one of the equal parts into which a scale is divided for convenience in tabulation. If we were tabulating the ages of employees, we should probably classify them by years. In this case the year would be the class interval. If we desired a more refined classification, we might classify their ages by half-years or by months. This is sometimes done in classifying the ages of children. If we were classifying people by their yearly salaries, we might select $100 as a convenient class interval. In

that case we should classify together all those who receive salaries between $1600 and $1699; in the next class interval would come all those whose salaries are between $1700 and $1799; and so on. We do likewise in classifying individuals according to their mental test scores. In *Figure 1* we have used ten as a class interval. For example, there are 21 students in the class interval 70–79.

The class limits. Let us suppose that a scale which runs from 0 to 100 has been divided into ten class intervals and that the intervals are designated as follows: 0–10, 10–20, 20–30, and so on. The tabulation would be easy except for the multiples of ten. In tabulating the number 30, for example, one would hesitate as to whether it belonged in the class interval 20–30 or in the class interval 30–40. In order to avoid this ambiguity in tabulation, it is customary to designate the class intervals so that they are mutually exclusive and do not overlap. There are three common ways of designating class intervals:

1. The class intervals in the above illustration may be designated as follows: 0–9.9; 10–19.9; 20–29.9; etc. The decimals are carried a little farther than the decimals that occur in the data. In this case the number 30 would be classified definitely in the class 30–39.9 and could not be classified in any other interval.

2. The same class intervals may be designated by the midpoint of each interval, thus: 5, 15, 25, 35, etc. In this case one tabulates each number in the class interval with the nearest midpoint. The num-

ber 27 would be classified in the interval of 25 because 27 is nearer 25 than 35. This method of tabulation should be avoided because it does not remove the ambiguity in tabulating multiples of 10.

3. The class intervals may be designated verbally as follows: 0 and less than 10; 10 and less than 20; 20 and less than 30; etc. This is unambiguous, but it is not so convenient on data sheets as the first method.

The *class frequency* is the number of cases (individuals or measures) in a class interval. According to *Figure 1* there are 25 students who received scores between 100 and 109, inclusive, in an intelligence test. The number 25 is therefore the class frequency for the class interval 100–109. It should be obvious that the sum of all the class frequencies is equal to the total number of cases in the table.

Problem 1. Each number in *Table 2* represents the score of a freshman at Lafayette College in an intelligence test. There are 253 students for whom we have records in this table. Prepare a frequency table for these scores similar to the frequency table in *Figure 1*. Supply the required information by inspecting the frequency table:

1. How many Lafayette freshmen obtained scores between 80 and 89, inclusive? (44)
2. How many students obtained scores between 110 and 139, inclusive? (43)
3. How many students obtained scores below 70? (61)
4. What percentage of the Lafayette freshman class obtained scores between 80 and 109, inclusive?
5. How many students in the Lafayette freshman class obtained scores of 110 or better?
6. What percentage of the freshman class at Lafayette obtained scores of 110 or better?

119	97	109	97	103	128	109	119	95	101	109	133	79	41	118
102	57	69	147	57	88	102	98	57	92	88	98	101	78	62
97	108	102	102	98	107	87	106	155	50	106	79	57	105	52
132	58	98	88	57	89	130	83	101	89	70	70	131	52	136
88	94	128	104	157	114	92	115	70	100	79	92	96	95	90
108	142	83	83	101	95	95	78	91	83	115	90	57	120	98
89	72	89	118	95	100	120	105	128	80	49	112	100	70	105
120	83	114	69	126	105	80	89	78	100	105	76	80	77	66
82	113	100	94	89	91	93	82	95	80	60	137	90	53	105
94	58	84	91	124	110	76	112	80	126	89	66	88	71	122
125	93	93	110	87	86	89	70	92	86	109	76	85	112	63
82	129	91	74	120	86	94	73	61	110	72	93	64	61	76
117	57	93	81	84	66	65	117	86	62	69	91	88	71	110
92	72	81	127	94	82	85	61	61	84	55	54	71	60	60
103	86	104	85	62	135	96	76	110	60	122	96	110	50	49
38	69	58	65	39	60	69	67	67	54	50	53	69	52	48
37	46	74	71	68	75	47	71	46	75	71	67	45		

Table 2. Scores of Lafayette College freshmen in an intelligence test

Chapter Two

The Column Diagram

When we are reading a report of some kind and come upon a page filled with numbers, we are tempted to skip them and confine ourselves to the text part of the report unless we have a technical interest in its detail. Few of us would stop long to examine *Table 1* in the preceding chapter because the significance of the data is not easily grasped. The data as arranged in the frequency table of *Figure 1* are much more intelligible. We can extract information from a frequency table more readily than from a heterogeneous list of scores. However, the magnitude of the several frequencies in a frequency table does not appear except by close inspection. The interpretation of a frequency table is made still easier by arranging the table in graphical form, and that is what we shall now do.

In a frequency table it is customary to arrange the scale with its class intervals in a vertical column, as shown in *Figure 1*. In plotting a diagram for a frequency table this scale should always be arranged horizontally, as is shown in *Figure 2*. At every class interval we draw a column, the height of which is proportional to the frequencies. The scale for the frequencies is at the left of the diagram. For example, we find from the frequency table in *Figure 1*

that there were 5 students who obtained scores between 50 and 59. This fact is also indicated in *Figure 2* by the 5 vertical units of the column at the class interval 50–59. Verify in the same way the heights of the other columns in *Figure 2* by referring to the frequency table in *Figure 1*. All the facts in

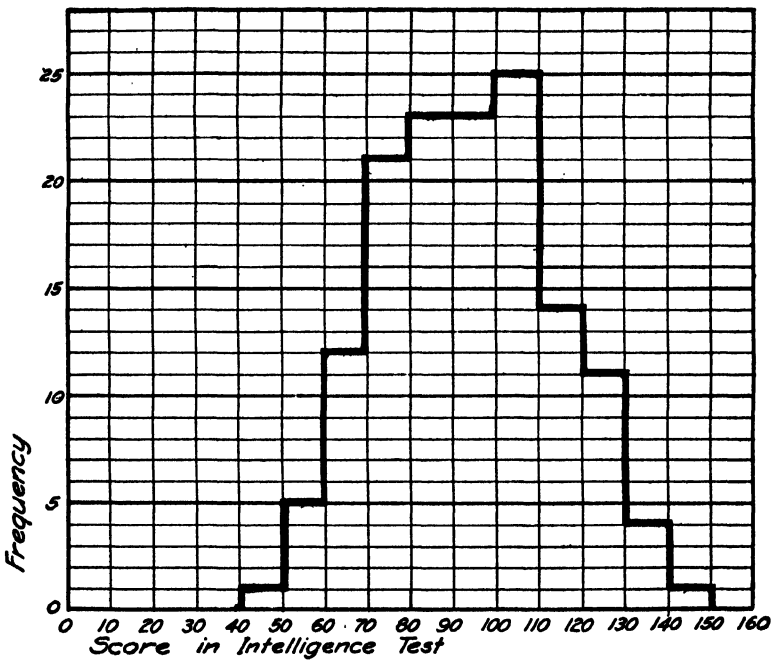


Figure 2. Column diagram with class interval of ten

the frequency table are represented in the *column diagram* of *Figure 2*, but the column diagram is preferable, especially when we want to compare several groups. The column diagram is sometimes called a *histogram*.

By reference to the column diagram such questions as the following may be answered:

1. Which class interval has the greatest frequency?

2. What is the approximate range? (Why is it that the exact range cannot be inferred from the column diagram?)

3. Which two adjacent class intervals have the same frequency?

4. Find two adjacent class intervals such that the frequency in one is roughly twice that of the other.

5. What is the frequency in the class interval 110–119?

6. How many students received scores between 90 and 109, inclusive?

The size of the class interval. This is largely a matter of convenience in representing the facts. In *Figure 2* we have used a class interval of 10, to correspond with the frequency table in *Figure 1*. If we arranged the same data in a column diagram with class intervals of 20 instead of 10, we should have a diagram like *Figure 3*. Consider, for example, the interval 40–59 in *Figure 3*. The height of the column is 6, which shows that there were 6 students who received scores between 40 and 59. Now verify the same fact by *Figure 2*. There we find 1 student in the interval 40–49 and 5 students in the interval 50–59, or a total of 6 students in the interval 40–59. Verify the same fact by reference to the frequency table in *Figure 1*. Compare *Figures 2* and *3* in the same way with reference to the class interval 60–79.

There are two points that should be kept in mind in determining the size of the class interval. These are: (1) the class interval should be small enough so that it will be possible to consider all the individuals in one class interval as identical for practical purposes;

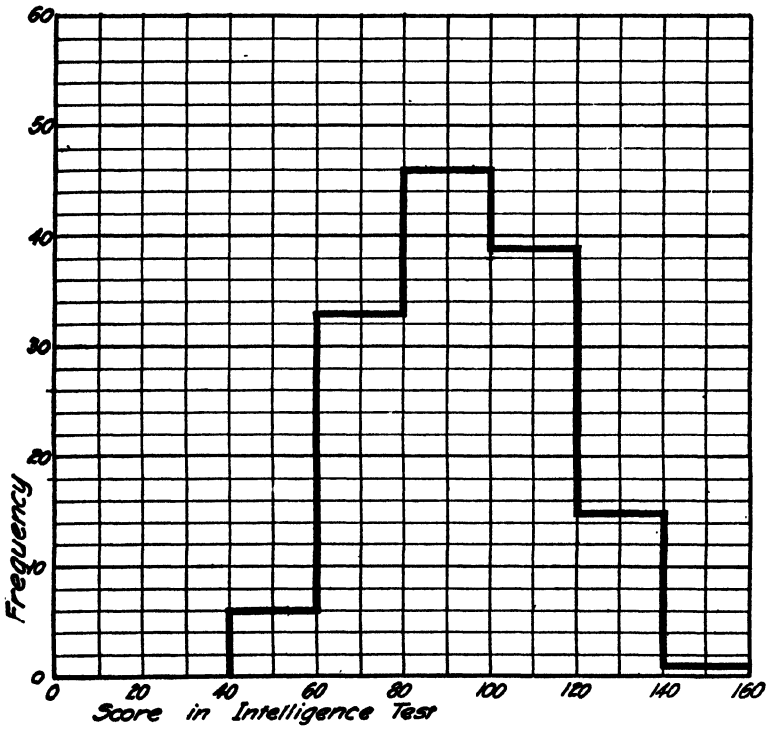


Figure 3. Column diagram with class interval of twenty

and (2) the interval should be large enough to avoid making the frequency table long and laborious to handle. In general, a variable can be handled to best advantage when its range is divided into from 15 to 25 class intervals.

The column diagram has an interesting property of which we shall make considerable use, namely, that *the area of a column diagram is proportional to the total number of individuals represented by the diagram.* Determine the area in *Figures 2* and *3* and verify the fact that both areas are 140, which is the

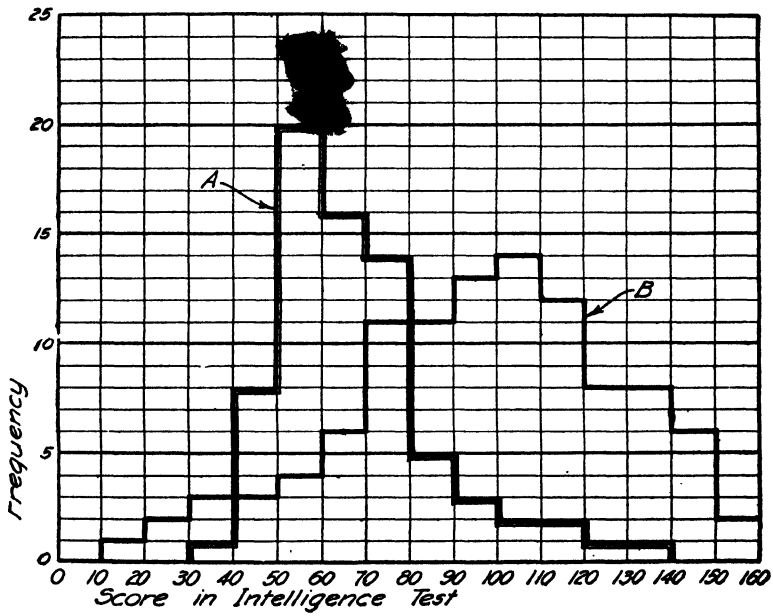


Figure 4. Superimposed column diagrams

total number of cases, as shown in *Table 1* and *Figure 1*. In doing this one must remember that the area represented by the column diagram is not necessarily the number of squares on the cross section paper that one happens to be using. Each class interval on the base line is counted as a unit. The units of the vertical dimension are the frequencies.

Problem 1. *Refer to the frequency table for Lafayette freshmen prepared in Chapter 1, Problem 1, and plot three column diagrams with class intervals of 10, 20, and 30 respectively. Determine the area of each of the three diagrams and verify the fact that the area is proportional to the number of cases. Label each diagram as shown in Figure 4.*

Problem 2. *Figure 4* represents the column diagrams for the intelligence test scores of two classes, A and B. Supply the required information by inspecting the diagrams in *Figure 4* and explain just how the diagrams give the information:

1. In which class is the brightest man as judged by the test?
2. In which class do you find the lowest man as judged by the test?
3. Which is the larger class? (Determine by inspection.)
4. Which class would probably have the higher average?
5. Which class has the larger range?
6. How many students are there in class A?
7. How many students are there in class B?
8. If the size of the class interval for a frequency table is increased, what happens to the number of class intervals in the table?

Chapter Three

The Frequency Polygon

The *frequency polygon* is a graph which is very similar to the column diagram and which tells the same story. It is, however, preferable because it shows more clearly than the column diagram the real nature of the distribution. There is one fundamental assumption in the graphic representation of frequencies which must be made clear at this point. In the column diagram of the preceding chapter we grouped the scores in classes. All the scores between 40 and 50 were grouped in one class, and the number of scores in that class was called the class frequency. For practical purposes we assume that all the scores in the class interval are at the midpoint of the class interval; namely, 45. In this way we are considering the people who get scores of 41 as equal to those who get scores of 49. Both are classified in the same class interval and considered as being at its midpoint; namely, 45. They are very nearly equal because the scale runs from 0 to 160; and if we divide the scale into intervals of 10, we divide the group into 16 classes, and this division is sufficiently refined for practical purposes. The assumption here is that all the scores in the same class interval are concentrated at the midpoint of the class interval.

The frequency polygon is constructed as follows (see *Figure 5*, which represents the same data as *Figure 2*):

1. Indicate the class intervals on the base line, using any suitable scale, as in constructing the column diagram.

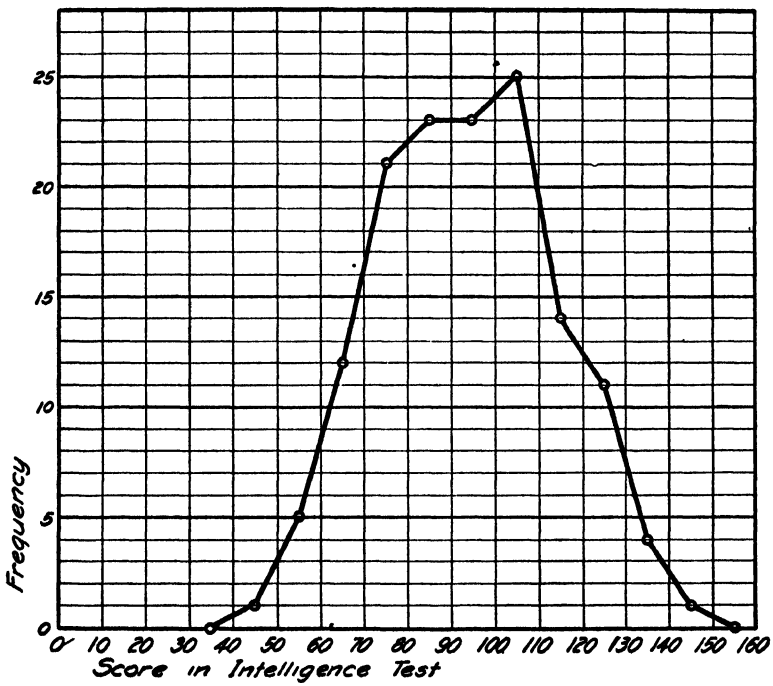


Figure 5. A frequency polygon

2. Indicate the frequencies on the left vertical margin of the diagram, as in constructing the column diagram.

3. Plot one point on the diagram for each class interval so that each point is right over the midpoint

of its class interval and so that its vertical distance from the base line will show the frequency of its class interval. The points of the frequency polygon are indicated by small circles in *Figure 5*.

4. Join these points by straight lines.

A frequency polygon can be constructed on top of a column diagram by joining with straight lines the midpoints of the horizontal lines for each class interval.

Compare *Figure 5* with *Figure 2* and satisfy yourself that the column diagram in *Figure 2* and the frequency polygon in *Figure 5* represent the same frequency table; namely, *Figure 1*.

Problem 1. Refer to the frequency table for Lafayette College freshmen in Chapter 1, Problem 1, and the column diagram for the same data in Chapter 2, Problem 1. Construct a frequency polygon for these data.

Chapter Four

Linear Relations

We shall now discuss the graphic representation of the relation between two variables. But we shall postpone the consideration of the corresponding equations for expressing such relations.

In *Figure 6* is indicated the customary arrangement for plotting a chart for two variables. The base line is called the *axis of abscissas* or *x-axis*. This is the axis that we have so far used for the scale of test scores. The left vertical margin of the chart is called the *axis of ordinates* or *y-axis*. This is the axis that we have so far been using to designate class frequencies. The scale on the x-axis is always plotted from left to right, as shown in *Figure 6*. The scale on the y-axis is always plotted running up, as shown in the diagram. Therefore both the x- and y-scales begin at the lower left corner. This point is therefore logically called the *origin*, and it represents zero for both scales.

It is poor form to present a chart without labeling both axes. One should also label them so that a reader may know what the units of the two axes represent, as *Temperature in degrees Celsius*, *Words remembered after one hour*, *Time in seconds*, and so on. The reader may then extract considerable information from your charts without consulting the text.

Where two variables are involved, it often happens that one of them is known and that we desire to know the second variable. The variable that is given or already known is called the *independent variable*, while

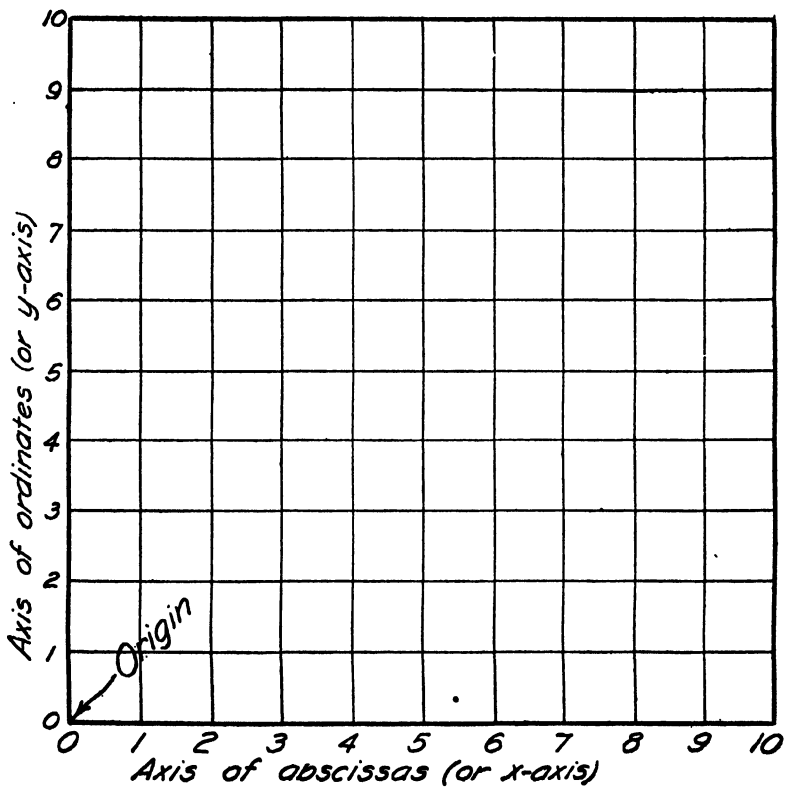


Figure 6. Abscissas and ordinates

the one that is to be derived is called the *dependent variable*. It is also customary to plot the independent variable on the x-axis and the dependent variable on the y-axis whenever this distinction can be made.

Suppose that you have a long list of numbers and that each number is to be divided by a constant, such as 17.6. A *constant* is a number that remains the same throughout a series of calculations. Variables are the varying quantities. If the list is long, we may dodge some of the labor of long division by doing it graphically. Let us use ordinary long division on four of the numbers and arrange them in a table. In this table we have called the given numbers x. In the second column we have placed the quotient $\frac{x}{17.6}$, which we call y. Verify this.

x	y
10	.57
34	1.93
53	3.01
97	5.51

Now we plot these four points as shown by the four small circles in *Figure 7*. The highest point, for example, is above 97 on the x-axis and directly opposite 5.51 on the y-axis. Verify the position of the other three points and compare these positions with the table above. The important fact is that these four points lie in a straight line. We draw the straight line through the four points, as shown.

Now we can find the quotients for the other numbers by inspection and save ourselves the labor of long division. For example, $\frac{78}{17.6}$ is 4.43. To get this we find 78 on the x-axis and run up to the line, where we find 4.43, which is therefore the

answer. The third figure is only approximate when we are using graphical methods. With a little practice one gains facility in reading charts of this kind. Find in the same way the quotients obtained by dividing 25, 39, and 87 by the constant 17.6. What

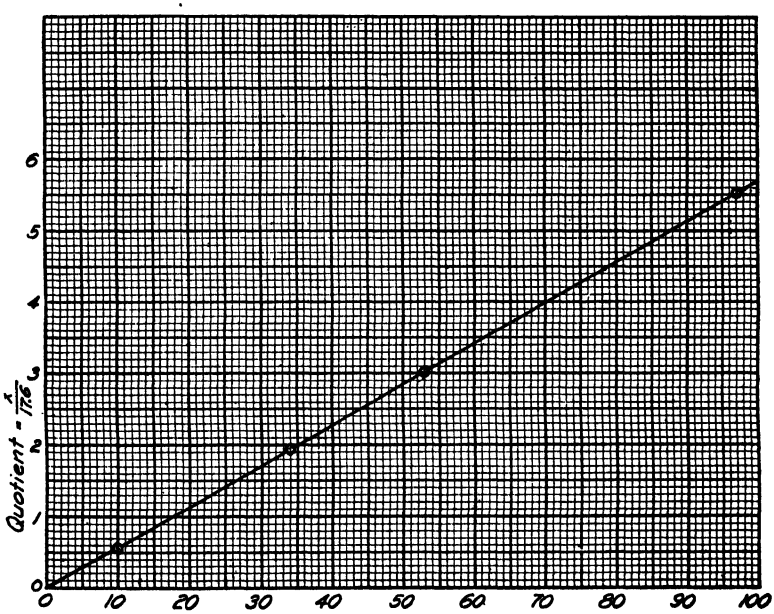


Figure 7. Graphical multiplication and division

is the quotient when 0 is divided by 17.6, according to the chart?

Let us plot the relation between inches and centimeters. In order to translate from one to the other, we use the formula:

$$1 \text{ in.} = 2.54 \text{ cm.}$$

Tabulating a few sample comparisons at random, we have the following table:

y IN.	x CM.
5	12.7
20	50.8
50	127.
70	177.8

Plotting these, we have the four points in *Figure 8*. These points lie in a straight line. Using the chart, we may say at a glance how many centimeters there are in 20 in., 36 in., 44.5 in., and so on. In the last illustration, 44.5 in., it is necessary to interpolate between the horizontal lines for 44 in. and 45 in. Facility in interpolating on charts is also readily acquired with a little practice. In the same way we may determine from the same chart how many inches there are in a given number of centimeters. Verify by inspecting the chart the fact that there are 39 inches in 100 centimeters.

It is highly essential that the student realize that in plotting lines on cross section paper he is not merely drawing lines on paper. Diagrams like *Figures 7* and *8* tell a story much more effectively and completely than whole chapters of text. It is highly essential that the student of statistics acquire the ability to see the meaning of a chart. There is probably not a single statistical idea which cannot be thought of more effectively in graphical form than in words. The student should constantly attempt to visualize statistical operations until this becomes a habit in thinking about any quantitative relations. By this I do

not mean mental arithmetic and visualizing numbers as such, but rather the lines and distances that represent them. Let me illustrate by a common example. If one is asked to find the square of .30, he probably thinks " nine hundred " as the answer, and then he thinks again, " Four decimals make .09." But we

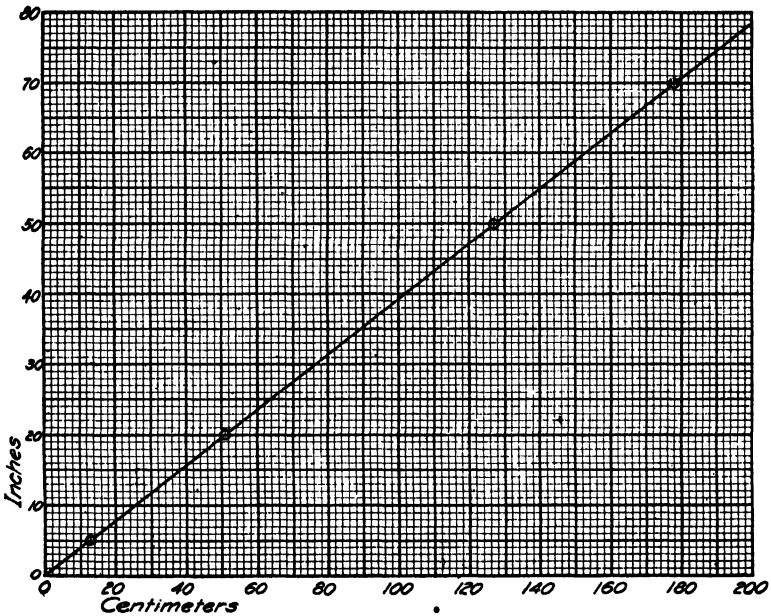


Figure 8. Graph for translating units of measurement

may not be sure that we are correct without counting the decimals with paper and pencil. Mental work of this kind is easier if one proceeds like this: Think of unity as a vertical line of any convenient height, in which .30 is about one-third of the line. The square of .30 is about one-third of .30, which is

approximately .10. In this way the reasonableness of calculating is made more apparent than by depending entirely on arithmetical operations. The square of .65 must be in the vicinity of .43 or .44 because .65 is about $\frac{2}{3}$ of unity, thought of as a unit distance. Two-thirds of .65 is somewhere in the vicinity of .44 because one-third of .65 is approximately .22. Numbers should be thought of as distances or other spatial magnitudes. It is more essential that the student learn to think of lines on charts, frequency polygons, and other graphic representations as alive, as actually moving to fit the facts, than to learn the arithmetical operations in statistics, no matter how blindly and conscientiously they are performed.

When it is necessary to deal with negative numbers on either or both of the two axes, the coördinates of *Figure 6* are extended as shown in *Figure 9*. If we compare *Figure 6* and *Figure 9*, we find that *Figure 6* is the upper right quarter of *Figure 9*. It takes care of positive numbers in both x and y and it is called the first quadrant, as shown in *Figure 9*. The other quadrants are added in order to make it possible to represent negative values on either or both of the axes.

The x-axis in *Figure 9* has a scale running from left to right, beginning with the lowest negative number on the axis. The y-axis runs up, beginning at the lowest negative number, as shown. The origin for a chart of this kind is at the center, as indicated. The origin is the intersection of the two zero-axes.

We shall now use the four quadrants for showing

the relation between two temperature scales, the Fahrenheit and the Celsius, or centigrade. By means of the chart we shall be able to tell at a glance what one thermometer would read when the reading on the

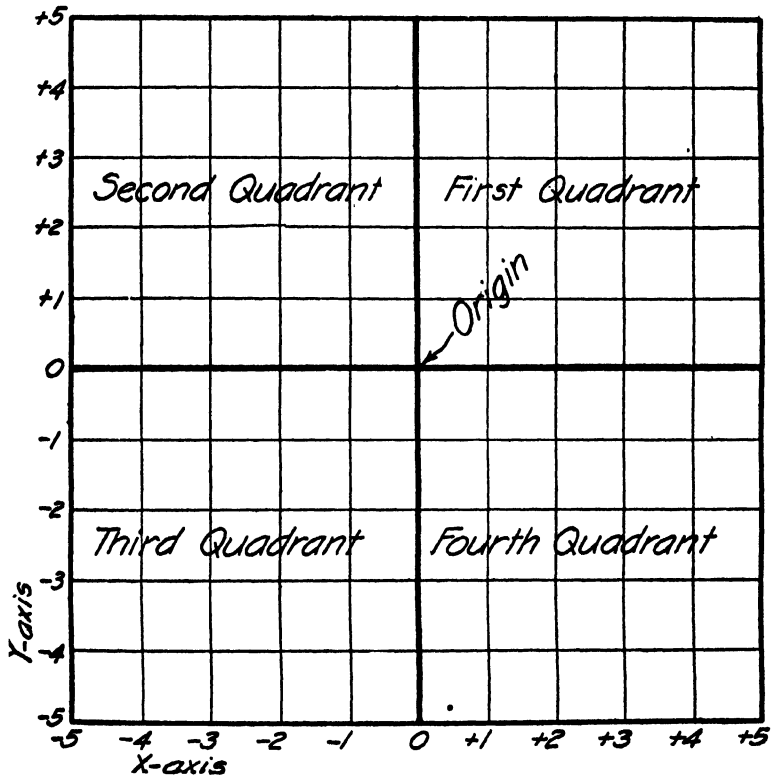


Figure 9. The four quadrants

other thermometer is known. First we get several paired observations at random, as in the table.

x Fahrenheit	68	−220	−76	199
y Celsius	20	−140	−60	93

These observations are transferred to a chart, as in *Figure 10.* Compare the four paired observations in the table with the corresponding points indicated by circles in *Figure 10.* It is now possible to translate from one temperature to the other without any calculations and merely by inspecting the diagram. For

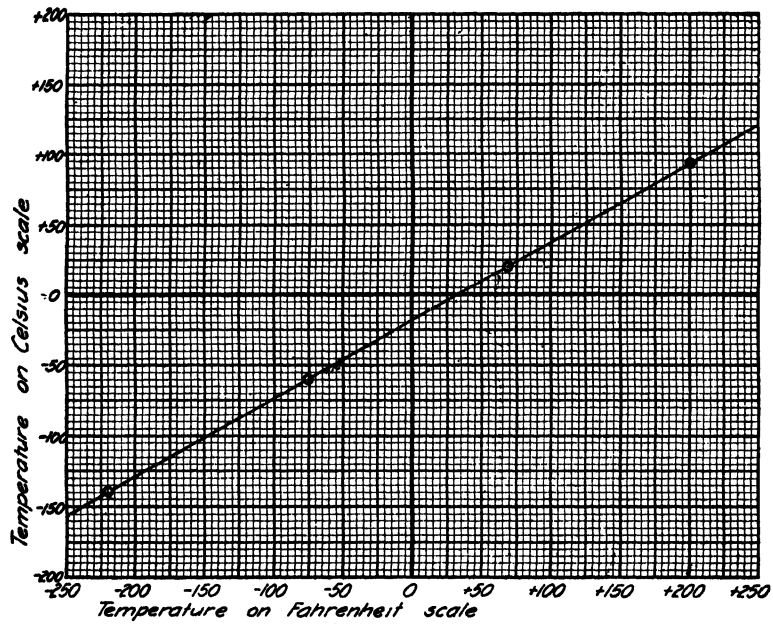


Figure 10. A relation involving both positive and negative numbers

example, when the Fahrenheit thermometer reads 90 degrees, a Celsius thermometer in the same room would read 30 degrees. Verify this on the chart.

The relations that we have just been plotting are all indicated by straight lines on the cross section paper. Such relations are called *linear relations.*

The straight line for a linear relation may or may not pass through the origin. If two variables, x and y, can both be zero at the same time, then the line passes through the origin. If the line does not pass through the origin, we can safely infer that when one of the variables is zero, the other cannot simultaneously be zero.

There are two other terms in connection with curve plotting which are useful. The reading on the x-axis at the point where a line or curve crosses the axis is called the *x-intercept.* In *Figure 10* the x-intercept is 32. It can be thought of as the value of x when y is 0. In *Figure 10* the interpretation of the x-intercept (32°) is that this is the temperature on the Fahrenheit scale when the Celsius scale reads zero. In the same way we define the *y-intercept* as the reading on the y-axis when x is zero. In *Figure 10* the y-intercept is approximately -18 degrees. It is the Celsius temperature when the Fahrenheit temperature reads zero.

Problem 1. Determine by inspection and without plotting in which quadrant each of the following points belongs:

Point	x	y
1.	$+3.2$	$+6.9$
2.	$+5.3$	-7.8
3.	-4.2	-1.7
4.	$+0.3$	$+2.1$
5.	-5.8	$+1.0$
6.	-3.4	0

Problem 2. On the following page is a series of paired observations. There are several errors in these observations. Chart them and discover the nature of the mistakes.

x	y
7.2	0.6
4.4	10.0
9.5	2.9
2.6	12.5
8.7	4.0
1.6	14.0
7.5	5.7
7.3	11.0
6.2	7.5
1.5	1.5
5.0	9.2
5.0	15.5

Problem 3. Supply the required information by inspection of *Figure 10:*

1. What is the Celsius reading when the temperature is − 68 degrees Fahrenheit?

2. If the Fahrenheit thermometer reads 212 degrees when water boils, what is the Celsius reading at the same temperature?

3. At what Fahrenheit temperature is the Celsius reading zero?

4. At what Celsius temperature is the Fahrenheit reading zero?

5. At what temperature are the two thermometer readings identical? $-5\frac{7}{9}^{\circ}$

Problem 4. Prepare a chart showing the relation between the diameter of a circle and its circumference. The length of the circumference is 3.14 times the diameter. Arrange the chart to show the relation between these two variables for diameters up to 5 in. Show that one can determine before plotting the chart that the line will pass through the origin. Arrange the scales to show dimensions in terms of $\frac{1}{4}$ in.

Problem 5. Prepare a chart to show the relation between the amount of time spent on a job and its cost. Labor for the job costs 80 cents an hour. There is a fixed charge of two dollars for each job for materials. Arrange the chart to show the cost of jobs up to four hours. Let the scale for time show intervals of fifteen minutes.

Chapter Five

Non-Linear Relations

We shall now discuss the construction and interpretation of charts in which the several points for sample observations fall in a curve instead of in a straight line. Such relations are called *non-linear.* As a matter of fact, it is customary to refer to this subject as *curve-plotting* even when the lines are straight. It is even customary to speak of a straight line on a chart as a curve, the word "curve" being then used to mean any kind of line on a chart.

Let us plot the relation between the side of a square and its area. First we tabulate a few sample observations like this:

x SIDE OF THE SQUARE	y AREA OF THE SQUARE
6	36
10	100
12	144
14	196

We plot the four points as indicated by the four small circles in *Figure 11.* The curve must pass through the origin because the square of zero is zero. This gives an additional point in plotting the curve. It is obvious that a curve is more accurate, the greater the number of points plotted. The points in *Figure 11* do not fall on a straight line, but in a curve. We

draw a smooth curve through the points, and we have a chart showing the relation between a number and

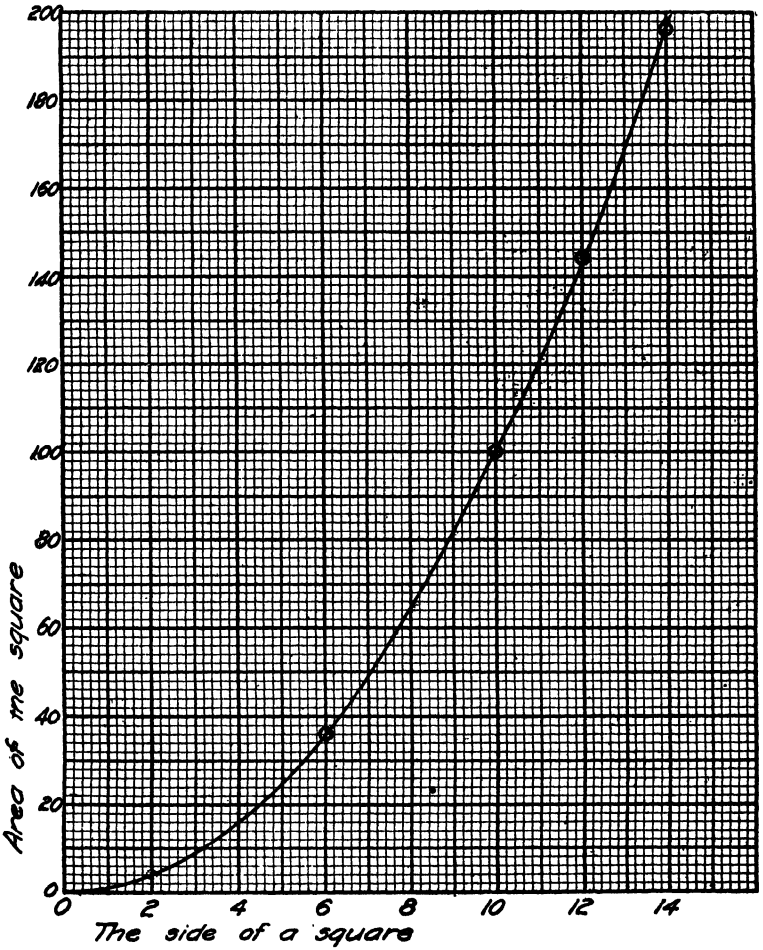


Figure 11. A non-linear relation

the square of that number. For example, if we wish to determine from the chart the square of 6.8, we

find 6.8 on the x-axis and then run up to the curve, which we cross at the level of 46 on the y-axis. This is as it should be because 46 is approximately the square of 6.8. One can also determine the square root of a number by a chart of this kind. For example, in order to find the square root of 78, we find

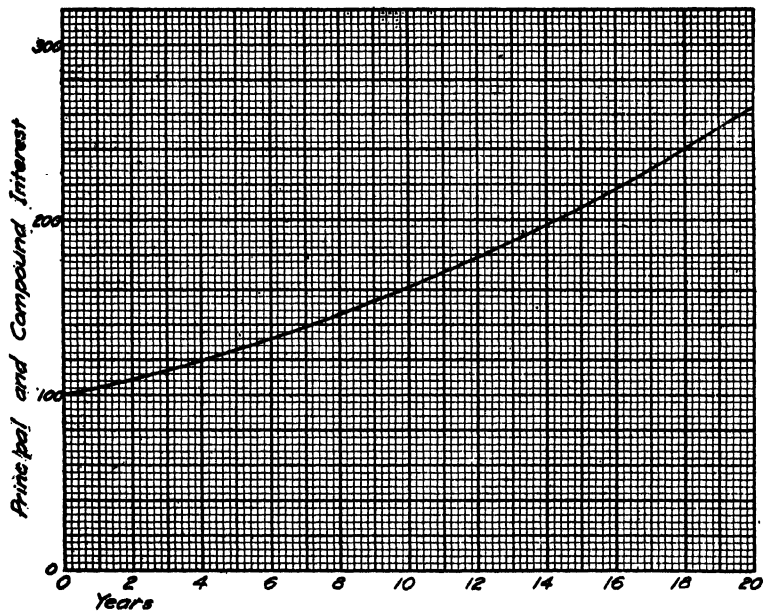


Figure 12. The curve for compound interest

78 on the y-scale and run over to the curve, which we cross at 8.8. This figure is approximately the square root of 78. A chart does not yield as accurate answers as does the arithmetical procedure. The chief advantage with the chart is that it shows the nature of the functional relation between the variables that

one is studying. In most scientific work that is much more important than decimal places.

In *Figure 12* we have another illustration of a relation that is non-linear. It shows how money grows at compound interest. The x-axis is time in years, and the y-axis is the total value of principal and accrued interest. The curve was plotted to show the increase of $100 when drawing interest at the rate of 5%, compounded annually. The chart shows, for example, that after 10 years the $100 have increased to a little over $160. One can also determine approximately how long the principal must draw the specified interest in order to reach a stated amount. Thus it will have doubled in a little over 14 years.

These facts can be obtained more accurately from an interest table, but the chart shows something that the interest table does not make quite so conspicuous. The chart shows that not only does the total amount increase but also that the rate of increase increases. The gain for the first 4 years is about $20, whereas the gain from 16 to 20 years is over $40. The curve not only rises but it rises more and more rapidly with time. This is the nature of the function, and such facts are not so conspicuous from a table of numbers.

Figure 13 should be studied in detail because it involves two fundamental points. It shows the progress of one person in the so-called mirror-drawing experiment, which is universally used in studying learning in psychological laboratories. The experiment consists in having the subject trace with a stylus the outline

of a six-cornered star, guided only by the reverse image of the star as seen in a mirror. This was repeated forty-two times in the experiment plotted in *Figure 13*. The first trials consumed more time than the subsequent trials. In fact, the subject's progress in learning the trick is measured by the time

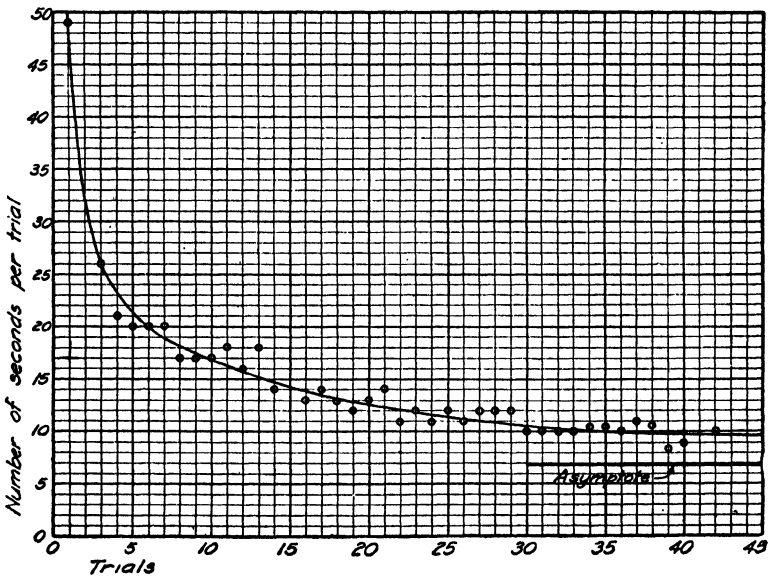


Figure 13. A curve to represent experimental observations

consumed per trial. The longer the practice, the shorter is the time per trial.

In *Figure 13* the x-axis represents number of trials. The y-axis represents time in seconds per trial. Each trial is indicated by a small circle. For example, the first trial consumed 49 seconds, as shown by the chart; the fifth trial required 20 seconds; the thirtieth trial required only 10 seconds.

When all the records have been plotted in the form of small circles, we may complete the chart in one of two ways. (1) We may join all the points by short straight lines, as we did in plotting the frequency polygon. That is sometimes done. (2) We may draw a smooth curve to represent the general progress of the subject in learning mirror drawing.

The location of a point on the chart is determined partly by the stage of proficiency attained by the subject and partly by many minor factors, such as distractions, amount of effort, state of fatigue, and so on. What we are interested in is a curve to show the relation between two variables; namely, practice as measured by the number of trials, and proficiency as measured by the time required to trace the outline. If we draw a smooth curve so that there are approximately as many points above the line as there are points below the line, we shall have a representation of progress as determined by practice. Such a smoothed curve will be relatively free from the minor disturbances of particular days. The smoothed curve in *Figure 13* was drawn by inspection.

There is a fundamental difference between the data of the social sciences and the data of the exact sciences as far as statistical treatment is concerned. This difference lies in the fact that in the social sciences our measures are not so refined as in the exact sciences. They are not relatively so free from chance disturbances as are observations in the exact sciences. For this reason we are compelled to dis-

cover trends and relations that may exist in spite of the many social chance disturbances over which we have no control. This is why we must establish tendencies by smoothing the curves so as to ignore the irrelevant factors. Practically all correlation statistics are based on the clear recognition of the fact that in the social sciences we are dealing with tendencies that may be discovered in spite of disturbing and irrelevant factors.

Having now plotted a curve to show the general relation between two variables, let us interpret this relation by inspecting the chart. It is at once apparent that the curve runs higher at the left side of the chart than on the right side. This means that the earlier trials in the mirror-drawing experiment require more time than the later trials, and that is reasonable because with practice one can reduce the time required to trace the star outline as seen in the mirror. Another fact which is apparent in the chart, but which does not appear so readily in the tables of raw data, is that the reduction of time in the first ten trials is greater than the time reduction in the last ten trials. More specifically, the subject reduced the time in the first ten trials from 49 seconds to 17 seconds, or a total time reduction of 32 seconds. In the ten trials from the 30th to the 40th the subject reduced his time by only 1 second. This is an example of the law of diminishing returns. Continued practice at almost anything yields improvement, but the first few hours of practice give more improvement than the last few hours. A relation of this kind is

seen in a chart at a glance, but it is not so readily seen in a table of numbers.

A learning curve like the one in *Figure 13* shows that the subject has not attained ultimate perfection in forty-two trials, and the chances are that he could never become so perfect that he would not show improvement with additional practice. This is indicated by the fact that the curve is still falling somewhat even after 42 trials. In order to ascertain how fast the subject could do the mirror drawing if the mirror coördination were completely mastered, he was asked to trace the outline, looking straight at it without the mirror interference. This required about seven seconds. It is reasonable to suppose therefore that, if he continued practicing indefinitely at mirror drawing, he would ultimately be able to trace the outline almost as fast with the mirror as without it; namely, in seven seconds. The heavy line at seven seconds may be considered a limit toward which the learning curve is approaching but which it will probably never quite reach. Such a straight line, indicating a limit toward which a curve is approaching but which it does not reach until the x- or y-variable becomes infinitely large, is called an *asymptote*.

Problem 1. Prepare a table of data for the x- and y-variables in *Figure 13* so that the table represents the observations in the laboratory from which the figure was charted.

Problem 2. The following table[1] enables one to determine

[1] These data are obtained from Yoakum and Yerkes' "Army Mental Tests," p. 133, Henry Holt & Co.

the score in the Army Alpha intelligence examination which corresponds to a given score in the Army Beta examination. However, in using the table it is sometimes necessary to interpolate for scores between the steps in the table. This is not necessary in a chart. Plot a chart showing the relation between these two variables and smooth the curve slightly if necessary.

Alpha	Beta
2	11
4	17
7	24
11	30
16	37
21	42
27	47
33	53
40	58
47	63
56	67
63	71
71	75
78	78
85	81
93	84
102	88
114	91
125	95
137	99
147	104
161	108

Chapter Six

Smoothing the Frequency Polygon

When a frequency polygon is plotted with a limited number of cases, the outline of the polygon is usually irregular. As the number of cases represented by the polygon increases, the polygon becomes smoother and shows more truly the real nature of the distribution. When a frequency polygon has been plotted, it is often desirable to smooth it so as to show what the polygon would be like with a larger number of cases. We cannot of course be absolutely certain that a smoothed polygon is exactly what it would be with a large number of cases, but we are quite sure that a smoothed polygon is more nearly like the polygon for a large number of cases than the irregular outline of a polygon for a limited sampling.

The chance deviation of a point on the frequency polygon above or below the true value may be inferred partly from the adjacent points. If a point on the polygon is below the points on either side, we may infer perhaps that its frequency is by some chance effect too low. If we inspect *Figure 5* for such irregularities, we find that the frequency of 25 in the class interval 100–110 seems abruptly too high, or it may be that the frequency of 23 in the adjacent class interval 90–100 is too low. If we gave the same mental test to the same class a second

time, we probably should not find the peak at 100–110, as in *Figure 5*. It is probably a chance fluctuation. In smoothing the frequency polygon we are trying to ascertain the shape which it would take if it represented conditions freed from minor accidental fluctuations. There are several slightly different methods of smoothing polygons. We shall use one

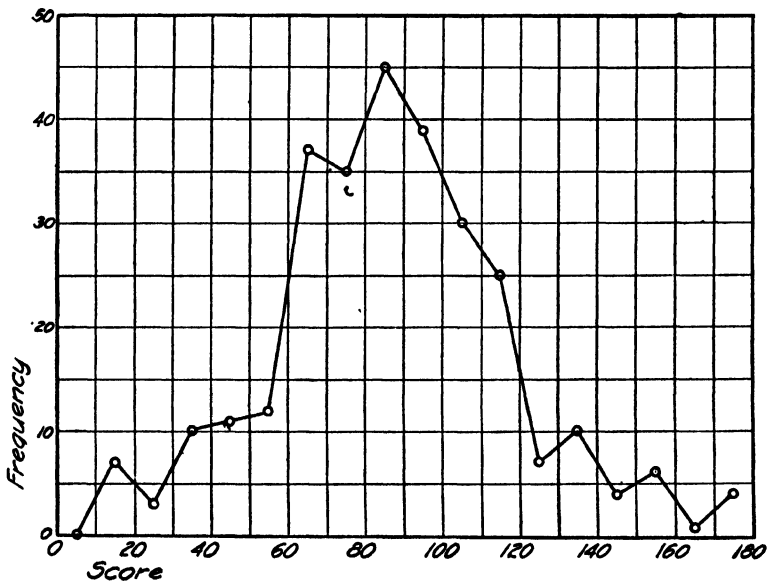


Figure 14. Frequency polygon before smoothing

method which can be handled either graphically or arithmetically.

Figure 14 shows a frequency polygon plotted directly from a frequency table. In *Figure 15* we have illustrated the procedure of smoothing the polygon, and in *Figure 16* we have the final smoothed

polygon as it would appear in a report. The distribution can be represented in a report either as in *Figure 14*, or as in *Figure 16*, or even in the more awkward form of a column diagram.

In *Figure 15* we have the frequency polygon represented by small circles joined by straight lines.

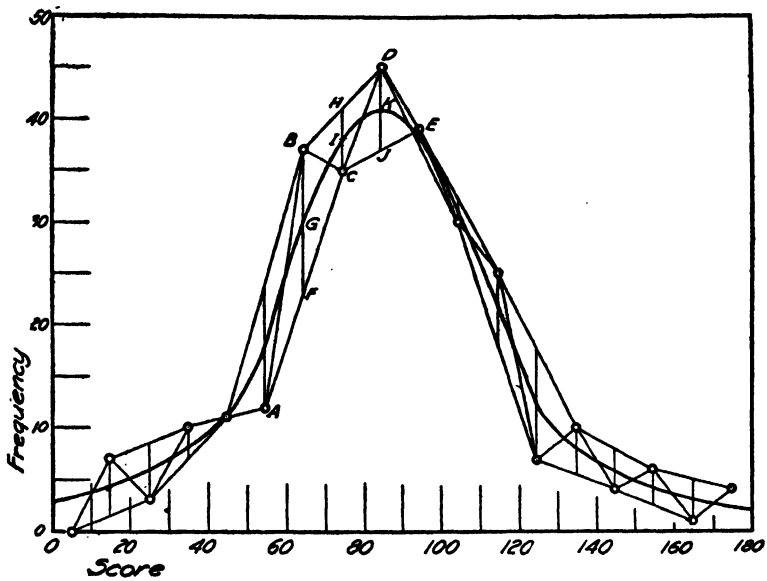


Figure 15. Frequency polygon with construction lines for smoothing

The cross rulings have been eliminated from this diagram in order to avoid confusion in studying the construction lines.

That part of *Figure 15* is identical with *Figure 14*. For convenience in explaining the procedure of smoothing, some of the points are labeled by letters. Notice the point C, for example. It represents a frequency of 35 for the class interval 70–79. It is lower than either of the two adjacent frequencies,

37 and 45, and we infer that perhaps it is slightly too low by some chance factor. In order to balance the frequency at *C*, we join the two adjacent frequencies by a straight line, *BD*. Mark the point, *H*, at which the line *BD* crosses the vertical through *C*. Notice the distance *HC* and locate by inspection the mid-point, *I*, of the line *HC*. This is 38, the balanced

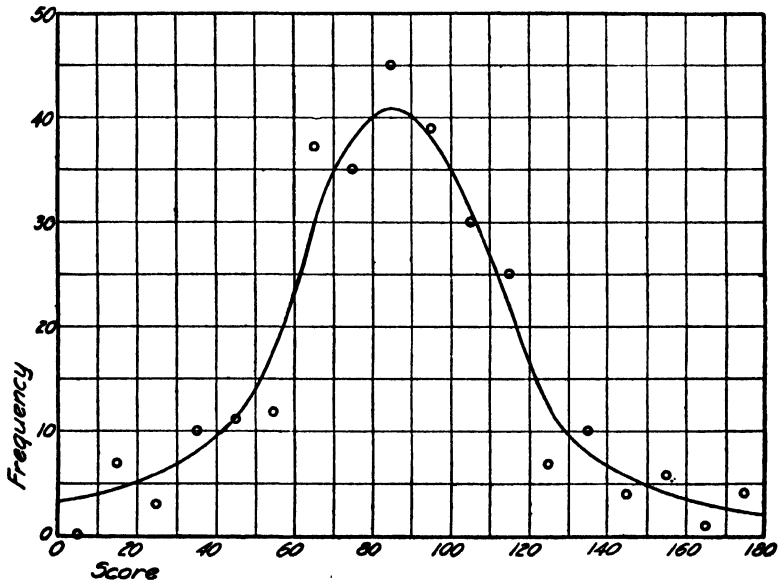


Figure 16. Smoothed frequency polygon with construction lines removed

frequency of *C*. In this way we have allowed the two adjacent frequencies to adjust the frequency at *C* so as to avoid the effects of small chance deviations as far as possible.

In order to balance the frequency of 45 at *D*, we join the adjacent frequencies by a straight line, *CE*,

which intersects the vertical through *D* at the point *J*. We locate by inspection the midpoint, *K*, of the line *DJ*. The balanced frequency for the class interval 80–89 is therefore 41. In the same way we locate the point *G*, which gives 30 as the balanced frequency for the class interval of 60–69, and continue likewise for the other points on the diagram. We then draw a smooth curve as nearly as possible through the balanced frequencies, as shown in *Figure 15*.

If we draw a smooth curve through the balanced frequencies or through the actually observed frequencies, such a line is called a frequency curve. The smoothed curves in *Figures 15* and *16* are called *frequency curves*, whereas if the frequencies were joined by straight lines, the outline would be called a *frequency polygon*.

In locating the balanced frequencies, one need not split hairs about the measurements. In fact one can safely draw the construction lines free-hand with a little practice. It is not necessary to go to conscientious extremes in smoothing a frequency polygon because the procedure is, even at best, an approximation of the outline that would be obtained with a larger number of cases.

In *Figure 16* we have transferred from *Figure 15* two things, namely, the frequency curve and the originally observed frequencies as indicated by the small circles. *Figure 16* shows the manner in which a frequency curve should be presented in a report. It is quite permissible, of course, to present a fre-

quency polygon in a report, and even the more awkward column diagram is occasionally used.

If a frequency curve is used in a report, be sure to show by small circles the actually observed frequencies of the original data so that the reader may use his own judgment as to the extent to which the original data have been smoothed. This is accomplished in *Figure 16*. The small circles show the original data, and the curve shows the author's interpretation of the nature of the distribution of which the data are a sample. This is a fundamental point to keep in mind in all curve plotting: the small circles show the actual observations, while the curve is an interpretation of the observations.

The arithmetical equivalent of the graphical procedure that we have just been discussing is as follows: In order to smooth the frequency of 35 at the point *C* in *Figure 15*, we add twice the frequency at *C* ($2 \times 35 = 70$) and the adjacent frequencies ($37 + 45$). This gives a total of 152, which we divide by 4 to get the balanced frequency of 38. These balanced frequencies should preferably not be tabulated because they might give the reader of a report the impression that they are observed frequencies, which they are not.

Another method of smoothing a frequency polygon is to increase the size of the class interval. When the size of the class interval is increased, there are naturally more cases in each interval, and that tends to smooth the effect of the minor fluctuations. This effect is seen more clearly when a frequency table is plotted with class intervals of varying size.

Problem 1. The following frequency table represents actual observations of the scores of a group of candidates in a mental test. Prepare three charts for these data as follows: one column diagram, one frequency polygon, and one frequency curve. On the frequency curve show the observed frequencies by means of small circles, as in *Figure 16*. Does the frequency curve have any asymptotes? Include in your report a diagram showing the construction lines for smoothing the frequency polygon.

Score	Frequency
22	1
21	1
20	1
19	6
18	16
17	13
16	31
15	47
14	52
13	62
12	70
11	62
10	67
9	57
8	58
7	41
6	42
5	29
4	12
3	11
2	4
1	5
0	2

Problem 2. The following is a frequency table for scores in a test given to a class of engineering students. Prepare three frequency polygons with class intervals of 1, 2, and 4 respectively. Plot the small circles over the midpoint of each class

interval in each chart. Remember that a score of 5 means five problems correctly solved and perhaps part of the sixth one. Plot all three charts with the same x- and y-scales.

Discuss the three polygons, comparing them as regards relative smoothness and the ease with which a curve could be fitted through the points, but do not draw the curves. This will show experimentally that increasing the size of the class interval makes the outline of the frequency polygon smoother and more continuous, and that the outline shows fewer zigzags.

Why is it that the three polygons are not of the same size when they represent the same number of cases and are plotted with the same x- and y-scales?

Score	Frequency
0	0
1	1
2	3
3	6
4	19
5	21
6	31
7	24
8	30
9	41
10	43
11	35
12	46
13	37
14	41
15	44
16	14
17	19
18	21
19	5
20	1
21	0
22	1

Chapter Seven

Graphical Tabulation

A frequency table is usually a means to an end. It is ordinarily a step in the preparation of a graph such as the frequency polygon. One can occasionally save considerable time in preparing a frequency polygon from raw data by plotting the original readings directly on the chart, thus eliminating the frequency table entirely. In *Figure 17* we have tabulated the data of *Table 1* and *Figure 1* directly, without the intermediate use of the corresponding frequency table. It is done as follows:

Read the numbers in the raw data, one at a time, and for each number place a dot above the appropriate x-value. The first number in *Table 1* is 62. Place a dot on the first line above 62. Never place these dots on the x-axis because it always represents zero frequency. The next number is 123. Place a dot on the first line immediately above 123 on the x-scale. Later in the series of numbers, 62 occurs again. When it is read off, we plot its dot on the second line above 62 because there already is a dot plotted on the first line above 62. When the whole list of numbers has been read off, we have the dots plotted as in *Figure 17*. We can now tell at a glance that there are five scores of 82 in the list, four scores of 102, none at 97, one at 96, and similarly for

other scores. The horizontal cross rulings enable one to read these frequencies directly, without counting the dots. In practice it is not necessary to make the dots as large as in *Figure 17*. In fact, pencil dots are sufficiently conspicuous on cross section paper unless it is printed in black.

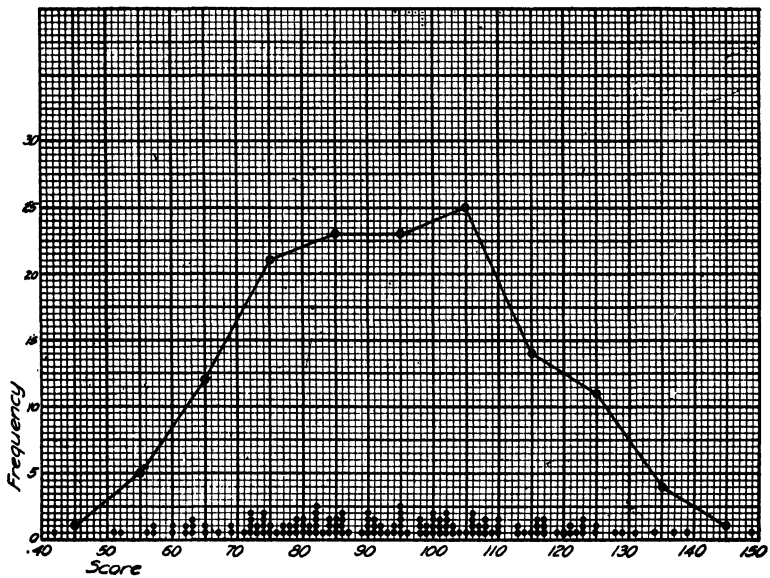


Figure 17. Graphical tabulation

When the dots have all been plotted, one should count them and check their number with the number of cases in the original data. The next step is to select a suitable class interval. *Figure 17* has been completed with a class interval of ten. The frequency scale for the polygon need not be the same as for the dots. To determine the frequency in any class interval, one simply counts the dots in that

interval. In *Figure 17* we count 25 dots for the interval 100–109. This class frequency is indicated by the small circle at an elevation of 25 over the midpoint for the class interval 100–109; and similarly for the other circles which are joined by straight lines to complete the frequency polygon.

A distinct advantage of this method of plotting is that one can readily replot the polygon with a different class interval if that is found desirable. This can be done without returning to the raw data. In plotting the dots one must be careful to keep his place in the table if he is interrupted.

50	64	70	68	64	72	65	55	79	55
75	48	60	51	55	46	51	66	62	58
47	72	47	65	74	67	55	49	46	65
51	58	63	54	60	62	82	61	73	50
51	68	83	70	54	63	54	77	51	70
77	50	65	64	59	66	65	51	55	63
46	48	79	67	82	72	57	65	58	72
66	62	58	68	52	58	59	78	66	48
74	73	53	61	62	73	67	60	48	64
46	57	60	77	78	53	51	55	68	49
72	50	52	59	58	60	68	63	53	57
59	83	67	65	55	59	51	60	61	58
44	51	78	64	62	50	57	67	69	55
66	68	61	57	68	61	59	50	60	56
48	57	65	54	59	65	76	64	54	64
60	67	58	66	63	54	63	60	62	51
56	48	73	60	57	73	52	56	58	47
78	64	52	52	56	58	68	52	77	56
54	59	63	65	67	63	60	58	46	60
56	60	54	61	57	50	66	49	47	49
54	54	75	46	60	74	58	72	56	43
70	60	44	56	63	45	56	60	44	63
42	46	64	61	67	40	62	64	71	37
63	69	43	33						

Table 3. Mental test scores of a class of students

Problem 1. The list of scores in *Table 3* represents the performance of a class of students in a test. Tabulate these scores graphically on cross section paper. Then complete the diagram in the form of a frequency polygon with class intervals of ten. From the graphical tabulation prepare a frequency table with class intervals of five.

Chapter Eight

The Equation of a Straight Line through the Origin

Almost every scientific investigation involves the discovery or study of the relationship between two or more variables. The relationship may be stated in different ways when quantitative relations are involved.

1. The relationship between two variables may be stated verbally. If we are studying the relation between the Fahrenheit and Celsius thermometers, we may describe this relation verbally as follows: "The Celsius thermometer reads zero degrees when water freezes, and it reads 100 degrees when water boils. The Fahrenheit thermometer reads 32 degrees when water freezes and 212 degrees when water boils. There is a constant ratio between a Celsius degree and a Fahrenheit degree." If we read 58 degrees on a Celsius thermometer and wish to translate this reading into a Fahrenheit scale, it is inconvenient to do so from this verbal description of the relation.

2. We may plot a graph for the relation as in *Figure 18*, in which the verbal description guided us in locating two points. These two points are connected by a straight line. Now, if we read 58 degrees on a Celsius thermometer and wish to translate it

into the Fahrenheit scale, we can readily do so by the chart. The corresponding Fahrenheit reading, as read off directly from *Figure 18*, is 137 degrees. The graph is in this case much more effective as a method of describing the relation between these two variables.

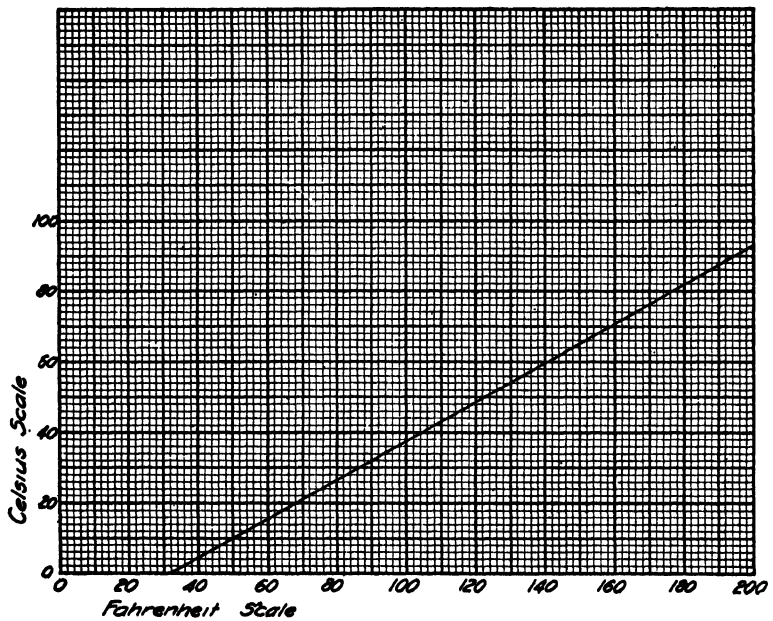


Figure 18. The graph of an equation

3. Another way to show the relation between two variables is by an equation. The equation for the relation between the two temperature scales is as follows:

$$C = .55\,F - 17.7,$$

in which C is the Celsius reading and F is the corresponding Fahrenheit reading. The derivation of

equations will be discussed shortly. If we read 58 degrees on the Celsius scale and wish to translate into the corresponding Fahrenheit scale, we simply substitute 58 for C in the above equation and solve for F thus:

$$C = .55F - 17.7$$
$$58 + 17.7 = .55F$$
$$\frac{58 + 17.7}{.55} = F$$
$$F = 137 +$$

We see, then, that the equation and the graph both tell the same story, but in different symbols. The graph can be read more quickly; the equation takes less space and is more accurate.

We shall consider first the equation of a straight line which passes through the origin. Its general form is

$$y = ax,$$

in which a is called the *multiplying constant.* In *Figure 19* we have charted seven equations of this kind. Notice, for example, the line for the equation $y = x$. This equation simply says that y is always equal to x, and that condition is true for any point on the line marked with this equation. When $x = +5$, for example, then $y = +5$. The point $(x = +5, y = +5)$ is on the line. The same condition is true for the extension of this line into the third quadrant.

Consider in the same way the line which is marked with the equation $y = .25x$. When $x = 6$, then

$y = 1.5$, according to the equation, and the point ($x = 6$, $y = 1.5$) is found on the line. In the same way one may easily verify to his own satisfaction the fact that each of the lines in *Figure 19* tells the same story as the corresponding equation. Substitute any number at random for either x or y and

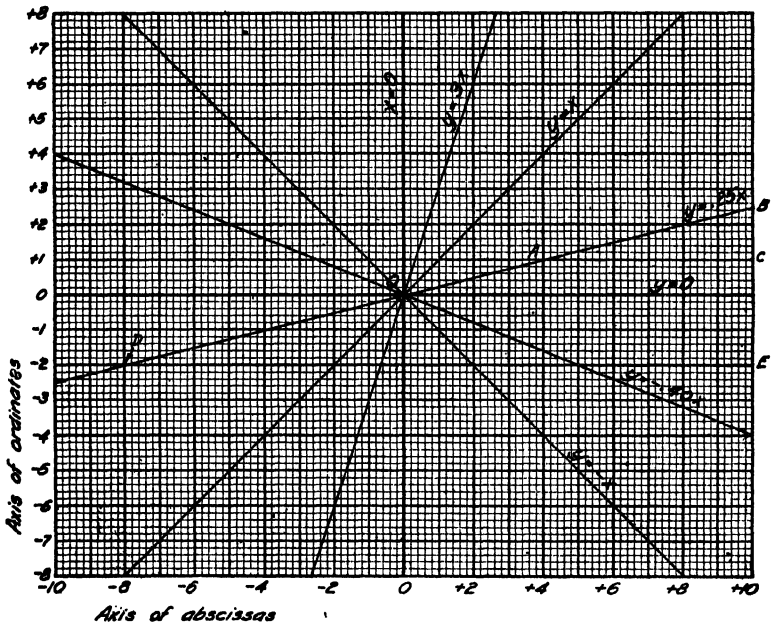


Figure 19. Straight lines through the origin, with their equations

solve for the other variable. The two values so paired will in each case be found on the line, or on an extension of the line if the chart were made larger.

The location of the lines in *Figure 19* is determined by the multiplying constant, for that is the only way in which these equations differ. *The multiplying*

constant is commonly known as the slope of the line. This is an appropriate name, for the greater the numerical value of this constant, the steeper is the line. Compare the three lines with the slopes of 3, 1, and .25. It is readily seen that the line with a slope of 3 is steeper than the line with a slope of only .25. The same observation about the slope can be made for the two lines in *Figure 19* that are drawn through the origin into the second and fourth quadrants. In these two lines the two variables x and y are opposite in sign.

The equation of the axis of abscissas or x-axis is $y = 0$, because on that line y is always 0. Similarly the axis of ordinates or y-axis has as its equation $x = 0$, because on that line all values of x are 0.

A line which satisfies an equation is called the locus of that equation. Every pair of x- and y-values that is satisfied by the equation can also be represented by a point which is somewhere on the locus of the equation. A pair of x- and y-values is said to be satisfied by an equation when the two members of the equation are equal with the x- and y-values substituted in them.

We have seen that a line can be plotted for a given equation. One can also determine the equation for a given line. If the line goes through the origin at 0, one can determine the corresponding equation by determining the slope of the line. That is done as follows: Sketch any right-angled triangle, as ABC, with the given line as hypotenuse (the long side of a right-angled triangle). Then one leg, AC, of the tri-

angle will be parallel to the x-axis. This triangle may be sketched anywhere along the line and may be of any convenient size. Measure BC and AC. The slope of the line is the ratio $\frac{BC}{AC}$. The distance BC is 1.5 and the distance AC is 6, as measured on the x- and y-scales. The ratio is therefore $\frac{1.5}{6}$ or .25, which is the slope of the line.

In order to increase the accuracy with which the slope is determined, it is best to make the triangle as large as possible. We could increase the size of the triangle by using the triangle DBE for our measurements. The slope is then $\frac{BE}{DE} = \frac{4.5}{18}$ or .25, as before.

We have now seen (1) that a line on a chart may be represented by an equation; (2) that an equation of the form $y = ax$ can always be represented by a straight line passing through the origin with a slope, or steepness, which is determined by the multiplying constant a; and (3) that an equation of the form $y = ax$ can be written for any line on a chart when the line passes through the origin. The graph and the equation are two ways of telling the same story.

Problem 1. Prepare a chart with four quadrants, as in *Figure 19*, and plot the following lines:

1. $y = +1.5x$
2. $x = -2y$
3. $x = \frac{3}{5}y$
4. $y = -.15x$

5. Plot the line that contains the two following points: $(x = +2, y = +6)$; $(x = 0, y = 0)$. Determine the equation of this line.

For each of these five equations assume any numerical value for one of the variables, solve for the other variable, and locate the point on the chart. Show that in each case the point so located falls on the appropriate locus.

Problem 2. Study *Figures 10* and *18* and show that these charts represent the same relationship although they differ in appearance.

Chapter Nine

The General Equation of a Straight Line

We shall now consider the general form of equation to describe any straight line on a chart. The general form is

$$y = ax + b,$$

in which a and b are constants while x and y are the two variables. We have previously seen that the constant a is called the multiplying constant and that it determines the slope or steepness of the line. The constant b is called the *additive constant*, and it determines where the line cuts the y-axis. In *Figure 20* we have five straight lines with their appropriate equations. Notice that all these lines are parallel. That could also be shown, before plotting the lines, by the fact that the multiplying constant, or slope a, is the same for all of them; namely, $+ .5$. Next notice that the additive constants are different. In fact, the additive constant is in each case equal to the y-intercept. The line with the equation $y = .5x + 4$ has as its additive constant $+ 4$ and this agrees with the fact that the y-intercept for that line is $+ 4$. Verify similarly the agreement between the additive constant and the y-intercept by examining the other lines in *Figure 20*.

By examining *Figure 20* we also discover that when the additive constant is positive, the line crosses the

y-axis above the origin, and that when this constant is negative, the line crosses the y-axis below the origin. When the line passes through the origin, the additive constant is zero, and the equation then takes the simpler form described in the previous chapter, as is

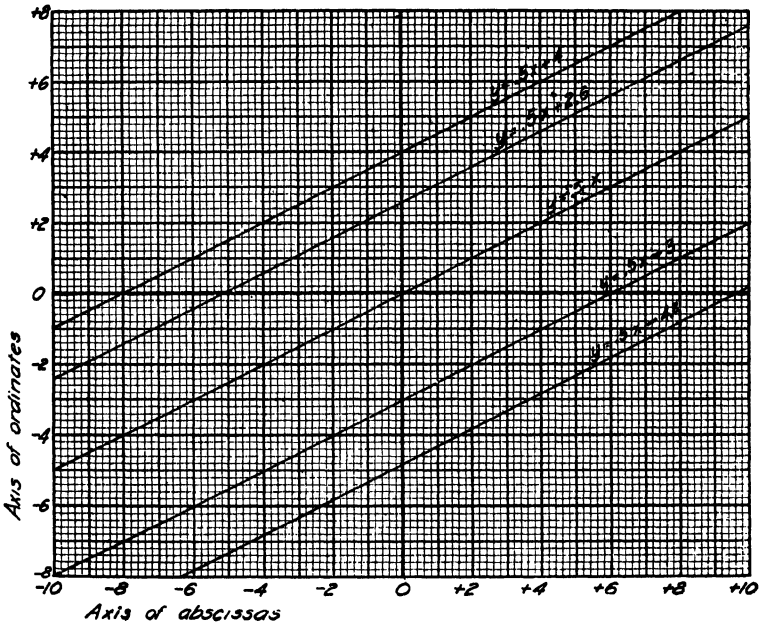


Figure 20. Parallel lines, with their equations

the case with the line in *Figure 20* showing the equation $y = .5\,x + 0$ or simply $y = .5\,x$.

In *Figure 21* we have drawn several straight lines at random. We shall now determine their equations by inspection of the chart. Consider first the line *A*. The general form of the equation for a straight line is $y = ax + b$. We must ascertain the numerical

value of each of the two constants in the equation in order to be able to use the equation. The additive constant b is simply the y-intercept. We find this on the chart at the point F, which reads + 2 on the y-scale. Hence the constant b is + 2. The slope

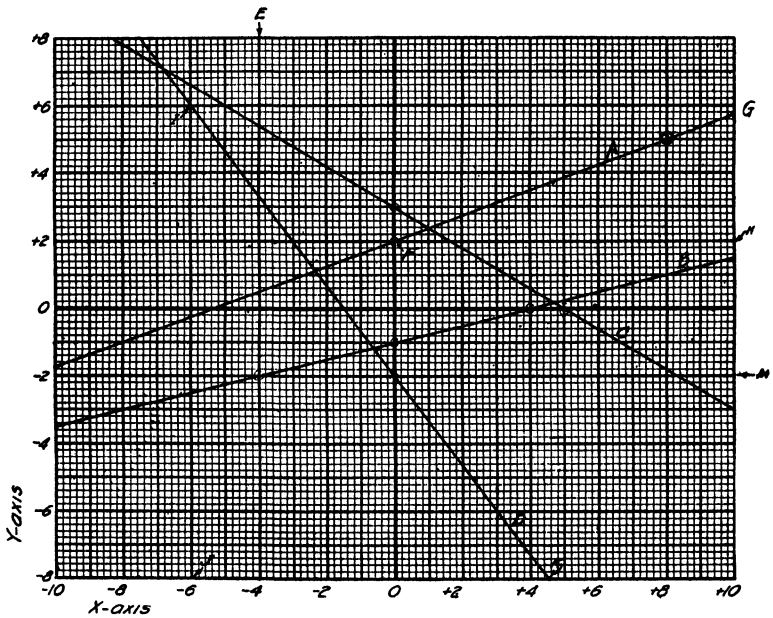


Figure 21. Straight lines, the equations of which may be written by inspection

a is determined as shown in the previous lesson. It is the ratio $\frac{GH}{FH}$, as read from the appropriate x- and y-scales; namely, $\frac{3.7}{10}$ or .37. The constant a is therefore .37. We can now write the equation of the line A thus:

$$y = .37x + 2$$

It is well to remember that when the line slopes down to the left, the slope is positive, and that when it slopes down to the right, it is negative. When the line is horizontal, as is the line M, the slope is zero and the x-term vanishes. The equation of the line M is therefore $y = 0\,x + b$ or simply $y = b$. The value of the additive constant b is again the y-intercept, and this is read directly from the chart as -2. Hence the equation of the line M is $y = -2$. This means that the value of y in the line M is always -2, which is readily seen to be true by inspecting the chart.

Referring to the equation for the line A, we may easily verify it by noting some point on the line, such as the point indicated by the small circle. Its coördinates are $(x = +8,\ y = +5)$. Substitute these values in the equation for line A thus:

$$y = .37\,x + 2$$
$$5 = .37 \times 8 + 2$$
$$5 = 2.96 + 2$$

The two members of the equation are as nearly equal as can be determined from a chart. If we do likewise with any point on the chart which is not on the line A, we discover that the two members of the equation are not equal.

The equation of the line E is written $x = -4$, which is another way of saying that the value of x for all points on the line E is -4. Inspection of line E, *Figure 21*, shows that this is actually the case.

We may determine in the same manner the equation for the line D. Its y-intercept is -2, and therefore the constant b is -2. The slope of the line D is found, as before, by the ratio of the two legs of any right-angled triangle against the line, such as IJK. The ratio here is $\frac{IJ}{JK}$, which is $\frac{14}{-10.5}$ or -1.33. The slope is therefore -1.33, and the equation of the line D is

$$y = -1.33x - 2$$

This equation can be verified, as before, by assuming any value whatever for x, such as -3. Substituting this value in the equation and solving for y, we find that $+2$ is the corresponding y-value. Locate this point, ($x = -3, y = +2$), on the chart and note that it falls right on the line D. In designating a point on a chart by its two coördinates, it is customary to write the x-value first and the y-value second. This makes it superfluous always to specify them as x and y. Thus the point ($x = -3, y = +2$) would ordinarily be designated simply ($-3, +2$).

In the same manner verify the fact that the equation of the line C is $y = -.6x + 3$. Keep in mind that the sign of a constant in an equation is an essential part of it. The student should form the habit of visualizing an equation of a straight line by noting two facts about each equation; namely, its slope and the y-intercept. It is sometimes useful to note that the ratio of the y-intercept to the x-intercept is equal to the slope.

We have discussed the methods of writing an equation for a given straight line. One can also plot a straight line on a graph to represent any equation which can be written in the form $y = ax + b$. In order to do this, simply assume two or three values for either variable and find by the equation the corresponding values of the other variable. Plot these points on a graph and join the points by a straight line. While only two points are necessary to locate a straight line, it is well in practice to plot at least three points in order to insure as far as possible against arithmetical error.

Occasionally we meet with an equation which at first sight does not look like the equation of a straight line, but which can readily be handled as such. We shall consider a few of these. The equation

$$y = 2x + 3 + 2$$

can readily be seen to be the equation of a straight line if we combine the two additive terms into one term, so that the equation becomes $y = 2x + 5$. Another example is $4y = 8x + 6$, which is the equation of a straight line and can be plotted by the methods that we have just described if we write it in the form $y = 2x + 1.5$. The equation $3x = 8 - 4y + 2x$ is also the equation of a straight line. We must express it, however, in such a way that the term in y is one member of the equation and the additive term and the term in x are the other member. When so stated, it is said to be written *explicitly in terms of* y. This is done as follows:

$$3x = 8 - 4y + 2x$$
$$x = 8 - 4y$$
$$4y = -x + 8$$
$$y = -.25x + 2$$

In the last form the equation is similar to the general form of equation for any straight line, in which the slope is $-.25$ and the y-intercept is $+2$.

Summary. The principles of the last two chapters may be summarized as follows:

1. Every straight line on a chart can be represented by an equation in the general form $y = ax + b$.

2. The slope or steepness of the line is indicated by the multiplying constant a. It is positive when the line slopes down to the left. It is negative when the line slopes down to the right. It is zero for all horizontal lines. It can be determined graphically by the ratio of the y-intercept to the x-intercept or by the legs of any other similar triangle on the chart. The measurements are always with reference to the x- and y-scales and not with reference to the rulings on the cross section paper that one happens to be using.

3. The additive constant indicates the y-intercept. It is positive if the line crosses the y-axis above the origin. It is negative if the line crosses the y-axis below the origin.

4. If two straight-line equations have the same multiplying constant, their loci must be parallel. If they have the same additive constant, they both cross the y-axis at the same point.

Problem 1. Draw on cross section paper four quadrants and plot the following equations;

1. $y = .6x + 3.4$
2. $y = 4.2x + 5$
3. $2.5x = -4.7 - 3y$
4. $14 - 9 + 2x - 1.8y = 0$
5. $(y - 1.6) + (x + .7) = 4$

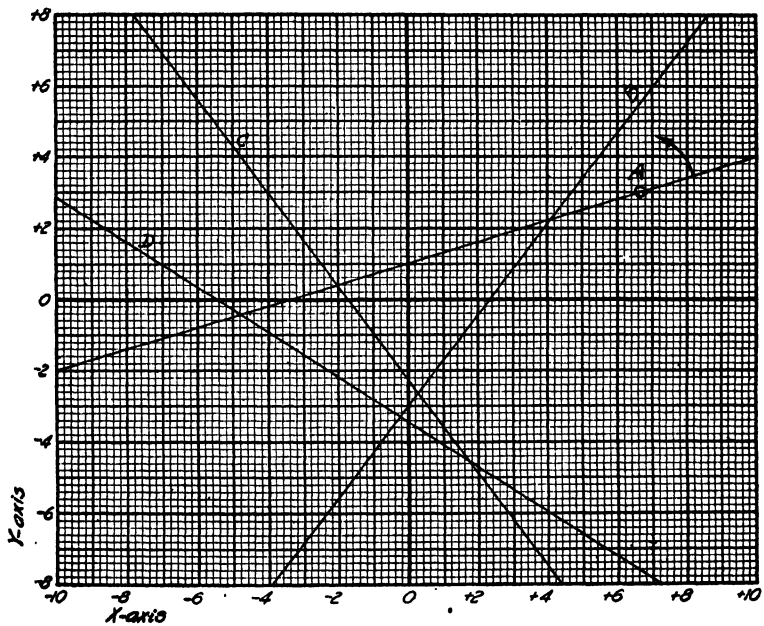


Figure 22

Restate each of these equations in the general form $y = ax + b$ before plotting them. After plotting the lines, assume any one point on the locus for each line, and show that the corresponding equation is satisfied by the numerical value of the coördinates for that point.

Solve equations 1 and 3 as simultaneous equations to determine the pair of values of x and y that are satisfied by both

equations. Then show that the solution represents the point on the graph at which these two lines intersect.

Problem 2. Copy the lines of *Figure 22* on cross section paper to any convenient scale and determine the equation for each line.

If the line *A* is swung around the point indicated by the circle in the direction of the arrow, what happens to the multiplying constant? What happens to the additive constant?

Under what conditions would a line with a multiplying constant of unity make an angle of 45 degrees with the x-axis?

Problem 3. Answer the following questions by inspecting the equations, but do not plot them:

1. Which of the following equations represent parallel lines?
2. Which line crosses the y-axis at the highest point?
3. Which line crosses the x-axis farthest to the left? Determine this by substituting zero for y and solving for the x-intercept mentally.
4. Which line passes through the origin?
5. Which of these lines slope down to the right?
6. Which of these lines would be the steepest if it were plotted?

1. $y = 4x + 3$
2. $y = 4x + 9$
3. $y = 2x + 0$
4. $y = -4x - 3$
5. $y = -2x + 0$

Chapter Ten

The Arithmetic Mean

If you had given an intelligence test to two classes and wished to determine which of the two classes was in general the brighter, you would probably calculate an ordinary average for each class. The class with the higher average would, in general, be the brighter class. The average is determined by adding all the scores and dividing the sum by the number of scores in the class. The quotient is the average. Stated more clearly and conveniently, we have

$$m = \frac{\Sigma S}{n},$$

in which m is the average of the scores, ΣS is the sum of all the scores, and n is the number of individuals in the group. This ordinary average is known in statistics as the *arithmetic mean* in order to distinguish it from other kinds of average; for there are several of them.

If we stopped to consider just what it is that we are doing when we calculate an average, we should conclude that we are trying to find a single number to represent a whole series of numbers, a single score to represent a long list of scores. It often happens that the average is a number which is different from any of the single numbers in the list that it represents.

For example, if we had scores of 2, 4, 6, and 8 in a small group of four, the average would be 5. This we should use as a single number to represent all four of the scores, even though the average, 5, is not represented in the original data.

Whenever a single number is used to represent a series of numbers, that number is called the *central tendency* of the series. The term is perhaps self-explanatory. The central tendency is a single number, or point on a scale, which is as far as possible the most representative of the series. It is usually a number somewhere in the central part of the range of the series. The arithmetic mean, or ordinary average, is the most generally used of all the measures of central tendency. We shall now consider four ways of calculating it. These differ only in detailed arrangement. Since they give practically the same average, the choice among them is simply a matter of preference as determined by the available calculating machines, the length of the series, the absolute size of the numbers, and the manner in which they happen to be classified. The arithmetic mean is the most commonly used mean and is usually referred to simply as the *mean*. There are other kinds of mean, such as the *harmonic mean* and the *geometric mean*, but these may be specified when necessary.

Calculation of the mean without tabulation. *Table 1* is a list of scores of students in a mental test. The mean score for that class can be figured perhaps most simply by adding all the scores on an adding machine. Divide the sum, 12,990, by 140, which is the number

of scores in the table. This gives a quotient of 92.785, which is the mean score. The student of statistics should learn to dodge arithmetical work, especially mental arithmetic, wherever machines are available for the drudgery. This leaves one's mind free to formulate the problem, which the machines cannot do.

Stated in convenient notation,

$$m = \frac{\Sigma X}{n},$$

in which m is the mean, X is any one of the numbers in the series, ΣX is the sum of all the X numbers, and n is the number of cases. The symbol Σ will occur repeatedly in subsequent work and should be clearly understood. It is not a separate quantity. It should be read "the sum of" what follows. Hence ΣX means the sum of all the X numbers. The above formula is therefore interpreted, "the mean is equal to the sum of all the X numbers divided by the number of cases."

Calculation of the mean from a frequency table. When the scores have been arranged in the form of a frequency table, as in *Table 4*, it is much more convenient to use the table for calculating the mean. One thereby saves the trouble of reading each of the original numbers. The procedure is to prepare a data sheet with headings as shown. In the first column are listed the class intervals. In the second column are listed the frequencies. In the third column we have the midpoints of the class intervals.

In the fourth column we enter the products of the items in the second and third columns. The mean is then determined by the relation

$$m = \frac{\Sigma f X_m}{n},$$

in which m is the mean, $\Sigma f X_m$ is the sum of the products in the fourth column, and n is the number of cases. It should be clearly understood that $\Sigma f X_m$ is the same as ΣX. This sometimes causes confusion in the minds of students. The notation $f X_m$ means the sum of all the X numbers in one class interval.

Scores	f	X_m	fX_m
40–49	1	45	45
50–59	5	55	275
60–69	12	65	780
70–79	21	75	1,575
80–89	23	85	1,955
90–99	23	95	2,185
100–109	25	105	2,625
110–119	14	115	1,610
120–129	11	125	1,375
130–139	4	135	540
140–149	1	145	145
	$n = 140$		$13{,}110 = \Sigma f X_m$

$$m = \frac{\Sigma f X_m}{n}$$

$$= \frac{13110}{140}$$

$$m = 93.64$$

Table 4. Calculation of the mean by a frequency table

It is the midpoint of the class interval multiplied by the number of cases in the interval. Hence $\Sigma f X_m$ means the sum of all the X numbers in all the class intervals. But that is what ΣX means also. The use of the notation $\Sigma f X$ instead of ΣX on data sheets is simply to make clear which columns are to be

multiplied, since the heading X is used also to designate the column which contains only the X-scale.

The mean, as determined by this method, is 93.64, which is practically the same as the absolute mean of 92.78, as previously figured for the same data.

Calculation of the mean by an equivalent scale. The products in the fourth column of *Table 4* are rather large, especially if it should be necessary to work with them without the aid of calculating machines. If the same result could be obtained with calculations involving smaller numbers, it would be an advantage. It happens quite frequently in statistical work that one can substitute a so-called equivalent scale, with smaller numbers than those of the original data, performing the calculations on this equivalent scale, and then applying a correction to the result in order to translate the equivalent scale back to the original data. Such a procedure sometimes saves considerable labor. We shall now calculate the mean of the same series of measures by the use of an equivalent scale (*Table 5*) so as to reduce the size of the numbers involved in the computations.

The mean expressed in terms of the equivalent scale, instead of the X-scale, is

$$m_e = \frac{\Sigma fE}{n} = \frac{681}{140} = 4.864$$

But we want this mean in terms of the original X-scale. This is accomplished by the relation

$$\begin{aligned} m &= 10\, m_e + 45 \\ &= 48.64 + 45 = 93.64 \end{aligned}$$

This transformation from the equivalent scale to the original X-scale involves two points of difference between the two scales. Inspection of the third and fourth columns of *Table 5* will show that the steps of the equivalent scale are very small compared with those of the midpoints, and that the equivalent scale begins with zero whereas the scale of midpoints begins with 45. Hence the above relation in changing from one to the other. The equivalent scale is of advantage only when the calculations involve large numbers which must be handled without calculating machines.

Scores	f	X_m	E	fE
40–49	1	45	0	0
50–59	5	55	1	5
60–69	12	65	2	24
70–79	21	75	3	63
80–89	23	85	4	92
90–99	23	95	5	115
100–109	25	105	6	150
110–119	14	115	7	98
120–129	11	125	8	88
130–139	4	135	9	36
140–149	1	145	10	10
	140			681

(1) $m_e = \dfrac{\Sigma fE}{n}$,

in which m_e is the mean in terms of the equivalent scale.

$m_e = \frac{681}{140} = 4.864$

(2) $c = I \times m_e$,

in which I is the number of X-scale units in each step of the equivalent scale.

$c = 10 \times 4.864 = 48.64$

(3) $m = c + m_a$,

in which m_a is the midpoint of the class interval of the X-scale which is designated zero on the equivalent scale.

$m = 48.64 + 45 = 93.64$

Table 5. Calculation of the mean by an equivalent scale

Calculation of the mean by an arbitrary origin. This is one of the most frequently used methods of calculating the mean, and it is also the method used for the mean in connection with correlation work, which will be discussed later. In using the equivalent scale we reduced the size of the numbers in the computation. This can be reduced still further by placing the zero of the equivalent scale somewhere in the middle range of the scale. The point on the X-scale at which we place the zero of the equivalent scale is called the *arbitrary origin* or the *assumed origin*. It can be placed anywhere at random, as far as accuracy of results is concerned, but it is most advantageously placed in the middle range of the X-scale because this reduces to a minimum the size of the numbers with which we deal in the computations.

In *Table 6* we have the detailed steps in the calculation of the mean by the method of an arbitrary origin. These data are identical with those of the two preceding calculations in this lesson. The computation is made as shown on the following page.

The first column in *Table 6* contains the X-scale arranged in class intervals, for convenience in tabulation. The second column contains the frequencies. In the third column we list an equivalent scale. This is done by first locating the assumed origin, zero, at one class interval somewhere in the middle range of the X-scale. It is advantageous to select the class interval to be designated as the assumed origin so that one has approximately as many cases in the f column above the assumed origin as there are cases

below it. This estimate need only be a rough one and is not essential to the arithmetical accuracy of the results. The steps of the equivalent scale are numbered consecutively + 1, + 2, + 3, etc. for the class intervals of the X-scale which represent numbers numerically higher than the assumed origin. The class intervals in the other direction are numbered consecutively on the equivalent scale − 1, − 2, − 3, etc. In *Table 6* we have placed the assumed origin at 95, which is the midpoint of the class interval 90–99. The class interval 100–109 is numerically the next higher and is therefore numbered + 1 on the equivalent scale. The class interval 80–89 is numerically the next lower one and is therefore numbered − 1 on the equivalent scale.

Scores	f	E	fE	
40–49	1	− 5	− 5	
50–59	5	− 4	− 20	
60–69	12	− 3	− 36	
70–79	21	− 2	− 42	
80–89	23	− 1	− 23	
90–99	23	0	− 126	$= \Sigma fE_{neg}$
100–109	25	+ 1	+ 25	
110–119	14	+ 2	+ 28	
120–129	11	+ 3	+ 33	
130–139	4	+ 4	+ 16	
140–149	1	+ 5	+ 5	
	140		+ 107	$= \Sigma fE_{pos}$

(1) $\Sigma fE_{pos} + \Sigma fE_{neg} = \Sigma fE = + 107 - 126 = - 19$

(2) $\dfrac{I \cdot \Sigma fE}{n} = c$

$\dfrac{- 19 \times 10}{140} = - 1.36$

(3) $m_a + c = m$

$95 - 1.36 = 93.64$

Table 6. Calculation of the mean by an assumed origin

The fourth column contains the products fE, *i.e.*, the products of the second and third columns, as shown. The sum of the positive fE products, ΣfE_{pos}, is + 107. The sum of the negative fE products, ΣfE_{neg}, is − 126. The number of cases, n, is the sum of the f-column.

The next step is to calculate ΣfE, which is simply the algebraic sum of ΣfE_{pos} and ΣfE_{neg}. In the example the sum ΣfE is − 19. This enables us to calculate the correction, c, by the relation

$$c = \frac{I \Sigma fE}{n},$$

in which c = correction. It is the difference between the true mean, m, and the assumed origin, m_a.

n = the number of cases.

I = the number of X-scale units in each class interval. It is 10 in the illustration.

When the correction, c, has been determined, we find the true mean by the relation

$$m = m_a + c.$$

In actually doing the computations for the sample problem at the end of this chapter it will be best to avoid this text matter and simply follow the three outlines in *Tables 4*, *5*, and *6*. These outlines are self-explanatory and should serve as a guide in computation without the text.

Before proceeding to calculate the mean by these various arithmetical procedures, it is well to stop and consider some of the properties of the arithmetic

mean. Suppose that a column diagram were plotted on heavy cardboard, or other substantial material, and that we cut the cardboard along its outlines. If we did this for the data we have been considering, we should have a piece of cardboard similar in appearance to the heavy outline in *Figure 2*. Now suppose further that we balance this piece of cardboard on a knife edge which is kept perpendicular to the base line. If we shifted the cardboard until it balanced, we should find that it would balance exactly at the absolute mean as determined in *Table 6*. This is a useful concept of the mean or ordinary average. It explains perhaps better than formulæ just what is meant by the mean.

All individuals are represented in the cardboard by equal rectangles. The width of each rectangle is the distance representing a class interval. The height of each such rectangle is the distance representing a frequency of one. We should keep in mind the fact that one individual located relatively far away from the knife edge causes the mean to " weigh " more, as it were, on that side of the knife edge. The tendency of each individual, as he is represented by a small rectangle in the column diagram, to overbalance the card on his side of the knife edge is called the *moment* of that individual. The moment is, strictly speaking, the product of the area or mass and its distance from the knife edge. Since all the small rectangles in the column diagram are equal, it is clear that the moment for any individual is determined solely by his distance from the mean. For this reason an

individual who is located at either extreme of the range has automatically more vote, as it were, in determining the mean. The individual who is right at the mean is entirely neutral. The practical point here involved is that a very few cases at high X-values or at low X-values affect the mean of the whole series very markedly, whereas a few changes in the frequencies of the class intervals near the mean do not noticeably affect the mean.

When a mean has been calculated, it is well to glance at the X-scale to see that the mean as calculated looks reasonable, *i.e.*, that it falls somewhere in the central range. If the mean as calculated falls outside the range, there is an arithmetical error somewhere.

Problem 1. Refer to *Table 3*. Calculate the mean for the data in the table by the four methods outlined in this chapter.

Chapter Eleven

The Median

The *median* is another measure of central tendency. It is defined as the middle number in a series of numbers when these have been arranged in their order of magnitude. The median score is the point on the X-scale so selected that there are as many scores above the median as there are scores below it. This condition is not true for the mean. Under some conditions the median is a better measure of central tendency than the mean.

Consider, for example, the following series:

7 14 11 2 17 1 22 13 9

Arranging these in the order of their magnitude, we have:

1 2 7 9 11 13 14 17 22

The median is 11 because it is the middle number when the numbers have been arranged in order of magnitude. There are four numbers in this series higher than the median, and there are four numbers below it. The mean is 10.7.

The above example has an odd number of cases. When the series has an even number of cases, the median is the number which is halfway between the two middle numbers. This satisfies the definition of the median.

Consider the following series:

11 29 8 4 10 3 17 37 22 7

Arranging these in order of magnitude, we have:

3 4 7 8 10 11 17 22 29 37

The two middle numbers are 10 and 11. The median is halfway between them, at 10.5. The mean is 14.8.

When the data are arranged in the form of a frequency table, it is advisable to calculate the median as shown in *Table 7*. The data of that table are identical with those of *Figure 1*.

X Scale	*f* Frequency	*Accumulated* Frequencies	
40–49	1	1	
50–59	5	6	
60–69	12	18	
70–79	21	39	
80–89	23	62	
90–99	23	85	8 cases below the median = *a* 15 cases above the median = *b*
100–109	25	110	
110–119	14	124	
120–129	11	135	
130–139	4	139	
140–149	1	140	
	$n = 140$		

$$\text{Median} = 90 + \frac{8 \times 10}{23} = 90 + 3.48 = 93.48$$

Table 7. Calculation of the median

The following steps indicate the method of calculating the median:

1. The first two columns are identical with those of previous frequency tables.

2. The third column contains the accumulated frequencies. For any class interval, such as 70–79,

the third column entry represents the sum of all the frequencies for this and the lower class intervals, *i.e.*, 1 + 5 + 12 + 21 = 39.

3. Ascertain the number of cases; in this case, 140. Determine $\frac{n}{2}$, which in the illustration is 70.

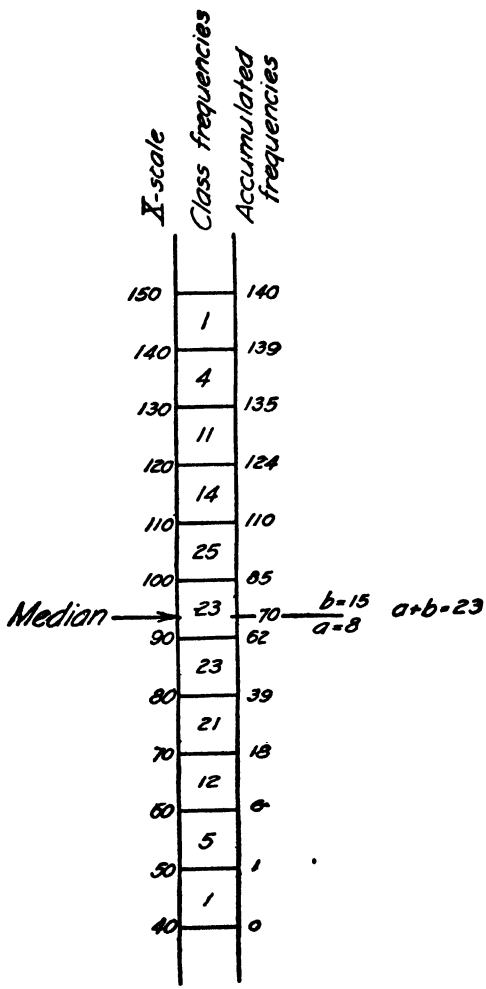
4. Find the class interval which contains the median. It is the class interval which, in the column of accumulated frequencies, contains $\frac{n}{2}$. In the illustration it is the interval 90–99.

5. Divide the class frequency of this interval into two parts, a and b, such that the accumulated frequency of the next lower interval plus a will be $\frac{n}{2}$. In the illustration this is 62 + 8 = 70. Hence $a = 8$; and since $a + b = 23$, b must be 15, as shown.

6. Let the number of X-scale units in each class interval be designated I. Then the median is the sum of the X-scale value at the lower edge of the interval which contains the median, plus $I\frac{a}{f}$, in which f is the frequency of that interval.

The calculation is very short and simple and is perhaps more readily understood by examining the computation at the bottom of *Table 7* than by the text.

In *Figure 23* we have illustrated further the calculation of the median. The median is somewhere in the interval 90–99. There are 23 cases in this interval. We assume that these 23 cases are distributed



$$\text{Median} = 90 + \frac{8 \times 10}{23} = 93.48$$

$$\text{Median} = 100 - \frac{15 \times 10}{23} = 93.48$$

Figure 23. The calculation of the median

uniformly through this interval. The median is the point on the X-scale so located that there are as many cases above it as there are cases below it. Starting at the bottom of the X-scale, we count off 70 scores and then note where we are on the X-scale. That point is the median. When we have counted 6 cases, we are at 60 on the X-scale. When we have counted 39 cases, we are at 80 on the X-scale. When we have counted 62 cases, we are at 90 on the scale. If we continued to 100 on the scale, we should pass 85 cases whereas we want to pass only 70; hence the median is somewhere in the class interval 90–99.

When we have counted the desired 70 cases, we have passed 8 out of the 23 cases in the interval 90–99. We assume that these 23 cases are uniformly distributed in the interval. Hence we go beyond the point 90 a distance equal to $\frac{8}{23}$ of the next interval. There are 10 scale units in each interval. Hence we add $\frac{10 \times 8}{23}$ to 90 in order to locate the median on the X-scale. The same reasoning can be followed by beginning at the top of the scale and counting down instead of up. Both procedures give the same median, as shown by the two sample computations in *Figure 23*.

It is of some interest to note that if on a column diagram we draw a vertical line through the median, we thereby divide the diagram into two equal areas.

Problem 1. Calculate the median for the data of *Table 2*. Check the median by calculating it from both the upper and the lower ends of the range.

Chapter Twelve

The Mode

The *mode* is a third measure of central tendency. It is simply the X-value at which the frequency curve is highest. When the data are arranged in the form of class intervals, with a class frequency specified for each interval, we assume the mode to be the midpoint of the class interval with the highest frequency. This assumption is roughly correct and serves the purpose in most cases. The mode therefore requires no calculating unless one cares to determine it by smoothing a frequency curve. This, however, is not ordinarily done.

We shall summarize briefly the definitions and properties of the three forms of central tendency.

The mean is the ordinary average. It is the sum of the scores in a series divided by the number of scores in that series. If a column diagram is plotted on cardboard and cut along the base line and along its outline, the diagram will balance on a knife edge placed at the mean.

The median is the middle score in a series of scores after the scores have been arranged in order from lowest to highest. The median score is such that there are as many scores above it as there are scores below it. Sometimes a novice gets the impression that the median is the middle of the range, but that is only rarely true and it is not a definition of the median.

The mode is the scale value at which the frequency curve is highest. It is the X-value at which the frequency curve has the highest ordinate. When the data are arranged in a column diagram, the mode is assumed to be at the midpoint of the class interval with the highest frequency.

It is well to remember that every measure of central tendency is a point on the X-scale. It sometimes happens that two or more factors influence the shape of a frequency curve so that the curve has two or more peaks. When a frequency curve has two peaks and when both of these are thought to be caused by factors other than chance, the curve is said to be *bi-modal*. To determine whether a curve is significantly bi-modal or whether the apparent bi-modality is simply caused by chance fluctuation in the class frequencies is largely a matter of statistical judgment.

When a frequency curve is asymmetrical, as in *Figure 24*, the curve is said to be *skewed*. When the frequencies in the class intervals on either side of the mode are well balanced, the curve is said to be *symmetrical*. The curve may be symmetrical without being normal. Inspection of the two curves in *Figure 24* shows that in curve *A* the right side has a longer sweep than the left side. In curve *B* we have the opposite condition. In order to differentiate them, we call these two kinds of skewness *positive* and *negative*. In curve *A* the long sweep is on the right side of the mode, toward the upper end of the X-scale, and it is therefore said to be *skewed positively*.

Curve *B* by the same kind of reasoning is said to be *skewed negatively.*

The three forms of central tendency fall at the same point on the *X*-scale if the distribution is symmetrical. If the frequency curve is skewed, the three forms of

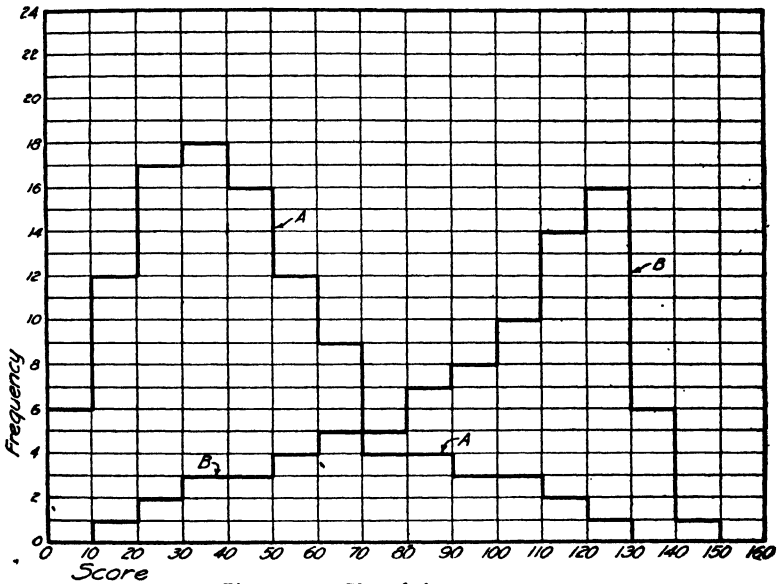


Figure 24. Skewed frequency curves

central tendency will fall at different points on the *X*-scale. We see here the reason for using several forms of central tendency. They become important when dealing with skewed surfaces.

Problem 1. Determine the mean, median, and mode for each of the two curves in *Figure 24.* Plot them on a chart and draw your own conclusions as to the manner in which these three forms of central tendency arrange themselves for skewed surfaces. Show that your conclusions are reasonable and consistent with the properties of the several forms of central tendency.

Chapter Thirteen

Variability

We shall now consider another fundamental concept in statistics. We have seen that the central tendency is a point on the scale which represents as well as possible the whole distribution. There are several measures of central tendency, and we have considered the three most commonly used; namely, the mean, the median, and the mode. If we desire to describe a distribution briefly, we specify its central tendency. For example, if we are dealing with the salaries of a group of men, we can give some idea of these salaries by saying that their central tendency, as expressed by the mean, is $3000. When we hear such a statement, we assume that some of the salaries are a little larger than $3000, that some of them are smaller, and that the average of all the salaries in the group is $3000. But, in order that a brief statement of the distribution shall be at all complete, it is necessary to state in addition how far the salaries scatter above and below the mean. For this purpose we state some measure of variability. In fact, a measure of variability shows how far the numbers of a series scatter on either side of the central tendency. If we are told the central tendency of a series of measures, and also to what extent the numbers scatter above and below the central tendency, we have in

these two facts a pretty fair idea of the size and distribution of the numbers. We are now concerned with the several methods of stating the variability of a series of numbers. In this and following chapters we shall consider four methods; namely,

1. The range
2. The mean deviation
3. The quartile deviation or semi-interquartile range
4. The standard deviation

The range. Let us compare two short series of numbers as to their central tendency and variability.

a.	5	10	15	20	25	30	35	40	45
b.	21	22	23	24	25	26	27	28	29

The mean of each series is 25; hence they are identical as far as the central tendency is concerned. The numbers in the first series scatter more than those of the second series. This is indicated by the fact that the range of the first series is $45 - 5 = 40$, whereas the range of the second series is only $29 - 21 = 8$. The range is one measure of variability. If we wished to describe these two series of numbers without enumerating them in detail, we could convey the information as follows:

	Central Tendency	Variability
Series a.	Mean = 25	Range = 40
Series b.	Mean = 25	Range = 8

A glance at these two facts about each series conveys a fair idea about them. We can see by the central tendency and variability of the two series that a num-

ber like 30 is above the average in each series, that it is relatively near the mean of series *a*, and that it is relatively very high in series *b*.

There is a serious practical limitation in the use of the range as a measure of variability which has made it imperative to use other measures for this attribute. The range is an unstable measure because it depends only on the two extreme cases, the highest and the lowest. If the series *a* and *b* represented mental test scores, the range for the whole group would depend on the two end cases only. The range is not affected by the variability of measures between the two extremes, as may be seen in the following illustration:

c. 5 22 23 24 25 26 27 28 45

Series *c* has the same range as series *a*, but its numbers do not scatter any more than those of series *b*, with the exception of the two isolated end cases.

The mean deviation. Let us consider another series of numbers such as this:

x, the series of numbers:	4	7	9	10	11	11	12	13	13	14	15	17	20
d, deviation from mean:	8	5	3	2	1	1	0	1	1	2	3	5	8

$\Sigma x = 156$ Mean of $x = 12$

$\Sigma d = 40$ Mean of $d = 3.08$

We have listed on the first line a series of numbers from 4 to 20, with a range of 16 and a mean of 12. On the second line we have listed the deviation of each number from the mean of the series, disregarding sign. Thus 15 has a deviation of 3 from the mean, which is 12; 20 has a deviation of 8; 9 has a

deviation of 3; and 12 has a deviation of 0. The sum $\Sigma x = 156$, and from this we get the mean, as $\frac{\Sigma x}{n} = 12$. The sum $\Sigma d = 40$, in which d represents deviation from the mean, and from this we get the mean deviation, as $\frac{\Sigma d}{n} = 3.08$. The mean deviation is simply the average of all the deviations, disregarding sign. If the mean deviation of a series of numbers is large, it shows that the numbers in the series scatter about the mean more than if the mean deviation were small.

The mean deviation can be calculated from any measure of central tendency, such as the mean or median. Hence one should always specify from which measure of central tendency a mean deviation has been calculated. In the above illustrations we have calculated the mean deviation from the mean of the series.

When a mean deviation is to be calculated for a series of numbers arranged in class intervals, we can arrange the calculation as shown in *Table 8*.

1. Tabulate the class intervals, the midpoints of the class intervals, and the frequencies in the first three columns, as in previous problems.
2. Add the frequency column to determine the number of cases, n.
3. Tabulate the fx products in the next column.
4. Add the fx column to determine Σfx.
5. Calculate the mean from the relation

$$\text{Mean} = \frac{\Sigma fx}{n}.$$

6. Tabulate in the next column the deviations, d, of the midpoints from the mean.
7. Tabulate the fd products.
8. Add the fd column to determine Σfd.
9. Calculate the mean deviation from the relation

$$\text{Mean deviation} = \frac{\Sigma fd}{n}$$

These calculations can be done to best advantage on calculating machines. If one must calculate a mean deviation for a large number of cases without the use of a calculating machine, the labor can be reduced by arranging an equivalent scale, as shown previously for the calculation of the mean.

We have seen previously that a measure of central tendency is a *point* on the scale. It should now be seen that a measure of variability is a *distance* measured in terms of the scale units.

Problem 1. Compare the three distributions of *Figure 25* as to their means and their mean deviations. Calculate the mean deviations from the mean of each distribution. In your report:

1. Show that it is possible to ascertain the mean of each of the distributions in *Figure 25* by inspection and without calculation. (Note that the three polygons are symmetrical.)

2. Show that one can determine by inspection and without calculation which of the two polygons, *A* and *B*, has the greater variability.

3. Show that one can determine by inspection and without calculation which of the two polygons, *B* and *C*, has the greater mean.

4. Arrange a table similar to *Table 8* for each of the three polygons *A*, *B*, and *C*. Read the class intervals and frequencies from the chart. Calculate the mean and the mean devia-

tion for each polygon shown in *Figure 25*. Compare the calculations with the answers to the first three questions.

5. State what happens to the mean and the mean deviation of the polygon *B* if it is shifted to the right without altering its shape.

6. Compare the variability of the three polygons by the range. Does this rank the polygons in the same order as the mean deviations?

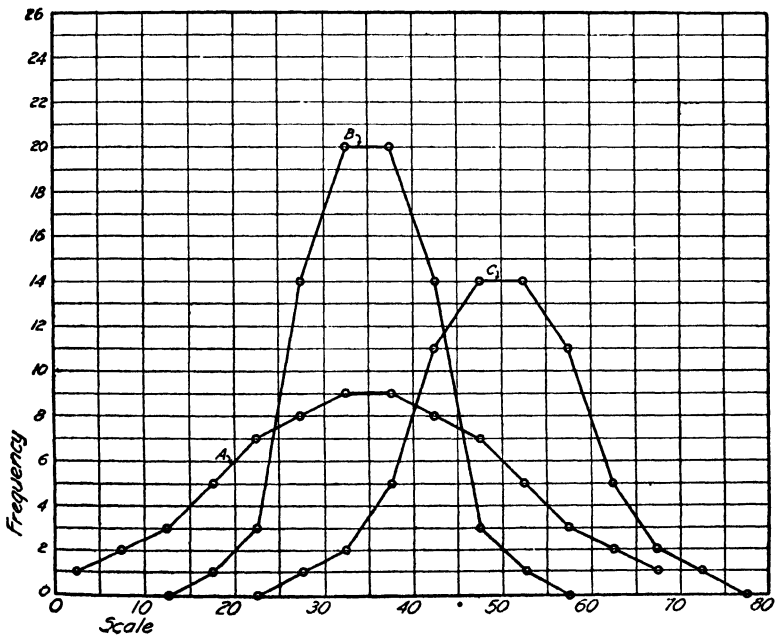


Figure 25. Three polygons showing differences in central tendency and variability

7. Suppose that the mean and mean deviation have been calculated for the stature of 500 men. What would happen to the mean and mean deviation if the calculations were repeated on 1000 cases?

Problem 2. Devise a method of calculating the mean deviation for the data in *Table 8* by the use of an equivalent scale with an arbitrary origin somewhere in the middle range of the

scale. Show the necessary formulæ. Adopt your own notation. Show that this procedure gives the same mean deviation as the calculations in *Table 8.*

Class Intervals	*Midpoint*	*Frequency*		*Deviation from Mean*	
	x	f	fx	d	fd
40–49	45	1	45	56.3	56.3
50–59	55	3	165	46.3	138.9
60–69	65	6	390	36.3	217.8
70–79	75	12	900	26.3	315.6
80–89	85	18	1,530	16.3	293.4
90–99	95	34	3,230	6.3	214.2
100–109	105	28	2,940	3.7	103.6
110–119	115	17	1,955	13.7	232.9
120–129	125	12	1,500	23.7	284.4
130–139	135	8	1,080	33.7	269.6
140–149	145	4	580	43.7	174.8
150–159	155	2	310	53.7	107.4
160–169	165	1	165	63.7	63.7
		$n = 146$	$\Sigma fx = 14{,}790$		$\Sigma fd = 2472.6$

Mean $= \frac{\Sigma fx}{n} = \frac{14790}{146} = 101.3$. Mean deviation $= \frac{\Sigma fd}{n} = \frac{2472.6}{146} = 16.94$.

Table 8. Calculation of the mean deviation

Chapter Fourteen

The Quartiles

As measures of variability we have so far discussed the range and the mean deviation. We shall now describe another measure of variability which is even more generally used, namely, the *quartile deviation* or *semi-interquartile range*. When we are examining a frequency distribution with special reference to its degree of scatter about the central tendency, the quickest way to ascertain it is to notice the range. If the range is large, the numbers scatter considerably on either side of the central tendency. The variability is then said to be great. We have called attention to the inadequacy of the range as a measure of variability, in that it is completely determined only by the two extreme cases, the highest and the lowest. Anything which causes a fluctuation in either or both of these extreme cases causes a fluctuation in the measure of variability of the whole distribution.

In *Figure 26* we have a column diagram of 68 cases. The median is at the point 11 because one-half of the area, 34 cases, is above the point 11 and one-half of the area is below it. The range for the diagram in *Figure 26* is 19. Instead of using this as a measure of variability, we shall determine the range which shows the scatter of the middle half of the group, *i.e.*, the middle 34 cases. For this purpose we count one-fourth of the area, 17 cases, up from the median and

locate the *upper quartile*, which is universally designated Q_3. The upper quartile in *Figure 26* is at the point 13 on the X-scale. Similarly we count one-fourth of the area, 17 cases, down from the median to locate the lower quartile at 8. It is universally designated Q_1. The median is sometimes, but not often, referred to by the notation Q_2.

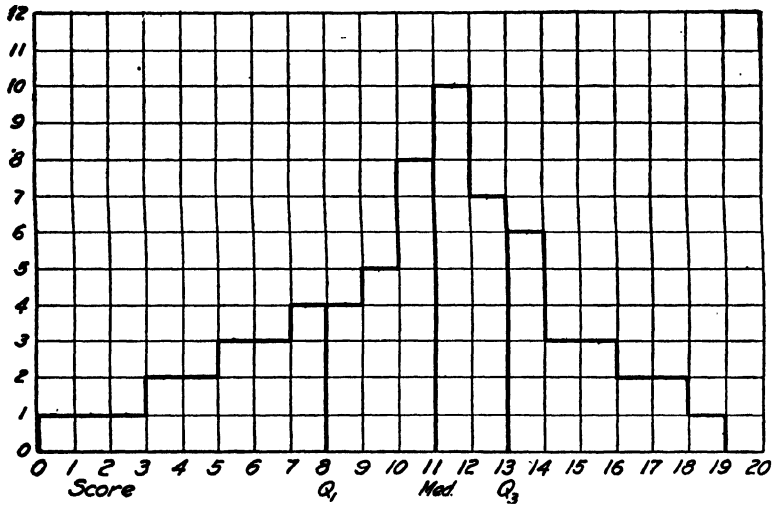


Figure 26. The quartile points

We have now divided the column diagram into four equal parts; namely, 0 – 8, 8 – 11, 11 – 13, 13 – 19. Each of these four parts contains 17 cases. The middle half of the numbers is between the lower quartile, Q_1, which is at the point 8, and the upper quartile, Q_3, which is at the point 13. This distance of 5 X-scale units between Q_1 and Q_3 is called the quartile range. It is a measure of variability which is quite generally used, easy to determine, and quite

stable as compared with the total range, which is unstable and unreliable. The quartile range is determined by all the measures and is therefore not so quickly affected by small fluctuations in one or two single numbers.

The range from the median to the lowest measure is 11. The range from the median to the highest measure is 19 − 11 or 8. Therefore the diagram is skewed low or negatively. This is also indicated by the fact that the lower quartile range (Median $- Q_1$) is greater than the upper quartile range (Q_3 − Median). A study of the location of the quartiles shows not only the variability or scatter of the numbers but also the relative degree of skewness and its direction.

Instead of using the quartile range as a measure of variability, it is quite common to specify the variability of a distribution by the quartile deviation, which is simply one-half of the quartile range. The quartile deviation of *Figure 26* is 2.5. The quartile deviation is sometimes called the semi-interquartile range.

If the distribution is symmetrical, not skewed, the upper and lower quartile ranges are identical; Q_1 and Q_3 are then equally distant from the median, and the median is in that case at the center of the total range. In such distributions the upper quartile range, the lower quartile range, and the quartile deviation are all identical.

We shall now calculate the quartile constants for another frequency distribution. The diagram in *Figure 26* was intentionally drawn so as to avoid decimals because they might confuse the explanation

of the quartiles. In practice the quartiles rarely fall at even integers, as in *Figure 26*. The method of calculating the quartile points is similar to the method previously described for the median. In fact, the median can be considered as one of the three quartile points which divide the distribution into four equal parts.

Procedure in calculating quartiles. Arrange a data sheet as shown in *Figure 27*. In the first column tabulate the class intervals. In the second column tabulate the class frequencies. Add the frequency column to obtain the total number of cases, n, which in the illustration is 143. Divide this by 4, which gives 35.75 cases in each quartile in the illustration.

Locate the class intervals which contain the three quartile points. This is done by counting 35.75 from either end for one quartile; 2×35.75, or 71.50 cases, for the median; and 3×35.75, or 107.25, cases for the other quartile.

When the class intervals containing the three quartile points have been located, divide the class frequencies in these class intervals so that the distribution is divided into four equal parts, as shown by the brackets in *Figure 27*.

Then locate the quartile point in the class interval, as shown by the calculations in *Figure 27*. These calculations are based on the assumption that the cases within a class interval are distributed uniformly through the interval. Thus the upper quartile point must be close to the upper end of the class interval 100–109 because its frequency, 25, is split so that

Class Intervals x-scale	Freq. f		
140 - 149	1		
130 - 139	5		35.75
120 - 129	11		
110 - 119	14		
100 - 109	25	4.75	
90 - 99	24	20.25 15.50	35.75
80 - 89	23	8.50	35.75
70 - 79	21	4.25 16.75	
60 - 69	12		
50 59	5		35.75
40 - 49	2		

Upper quartile, $Q_3 = 110 - \frac{4.75 \times 10}{25} = 110 - 1.9 = 108.1$

Median, $Q_2 = 100 - \frac{15.5 \times 10}{24} = 100 - 6.5 = 93.5$

Lower quartile, $Q_1 = 80 - \frac{4.25 \times 10}{21} = 80 - 2.0 = 78.0$

Figure 27. Calculation of quartiles

only 4.25 cases belong to the top quarter while the remaining 20.25 cases belong to the next lower quarter. We therefore place the upper quartile point $\frac{4.75}{25}$ of the distance of a class interval from the top of that interval. Hence we get:

$$\text{Upper quartile} = 110 - \frac{4.75 \times 10}{25} = 110 - 1.9 = 108.1$$

The other quartile points are located by the same kind of reasoning.

Having located the three quartile points, we determine the quartile constants directly from their definitions, as shown in the following calculations:

$$\text{Quartile range} = Q_3 - Q_1 = 108.1 - 78.0 = 30.1$$

Quartile deviation, or semi-interquartile range,

$$= \frac{Q_3 - Q_1}{2} = \frac{108.1 - 78.0}{2} = 15.05$$

$$\text{Upper quartile range} = Q_3 - Q_2 = 108.1 - 93.5 = 14.6$$

$$\text{Lower quartile range} = Q_2 - Q_1 = 93.5 - 78.0 = 15.5$$

We might summarize the properties of quartiles as follows:

1. The three quartile points divide the distribution into four equal parts. Each quartile has the same number of cases; namely, one-fourth of the total distribution. The four areas of the column diagram marked off by the three quartile points are equal. See *Figure 26.*

2. If all the numbers of the distribution were written on separate slips of paper and thrown into a hat, the chances would be even that any number

drawn from the hat would fall between the upper and lower quartiles. This is reasonable because one-half of the numbers are between the upper and lower quartile points while the other half of the numbers are above or below these points.

3. Any one of the quartile constants gives a fair idea of the degree of scatter or concentration of the numbers about the central tendency.

Problem 1. The following table shows the frequency distributions of two groups of people in an intelligence test. These groups are designated *A* and *B*. Calculate for each group the median, upper quartile range, lower quartile range, and quartile deviation, and make a comparison of the two groups of people on the basis of these constants. Plot both distributions in the form of frequency polygons on the same chart and indicate on it the quartile constants.

Class Intervals	*f* Group A	*f* Group B
20–29	0	0
30–39	2	0
40–49	4	2
50–59	4	4
60–69	6	8
70–79	10	12
80–89	16	16
90–99	14	8
100–109	10	6
110–119	8	4
120–129	2	4
130–139	2	4
140–149	0	4
150–159	0	2
160–169	0	2
170–179	0	2
180–189	0	0

Chapter Fifteen

The Standard Deviation

We shall now discuss one of the most generally used measures of variability; namely, the *standard deviation.* We have already seen that the mean deviation is the mean of all the deviations, disregarding sign. The standard deviation is of the same order of magnitude. It differs from the mean deviation simply in this: the deviations are squared before being summed; this sum is divided by the number of cases, as with the mean deviation; and out of this quotient we extract the square root, in order to obtain the standard deviation. This last step is necessary, from one point of view, in order to reduce the measure of variability to the same order of magnitude as the original deviations.

The standard deviation is universally symbolized by the notation σ (sigma), and it is expressed more briefly this way:

$$\sigma = \sqrt{\frac{\Sigma d^2}{n}},$$

in which d represents deviations from the mean of the series, and Σd^2 represents the sum of the squares of all deviations. The advantage of the standard deviation is that it can be handled algebraically better than the other measures of variability.

In *Figure 28* we have a frequency curve, with its mean at 100 and a standard deviation of 30. This means that about two-thirds of the cases are located between 70 and 130. The standard deviation is a distance, as all measures of variability must be. In *Figure 28* we have plotted both the original X-scale and the corresponding sigmas. The number 70 can

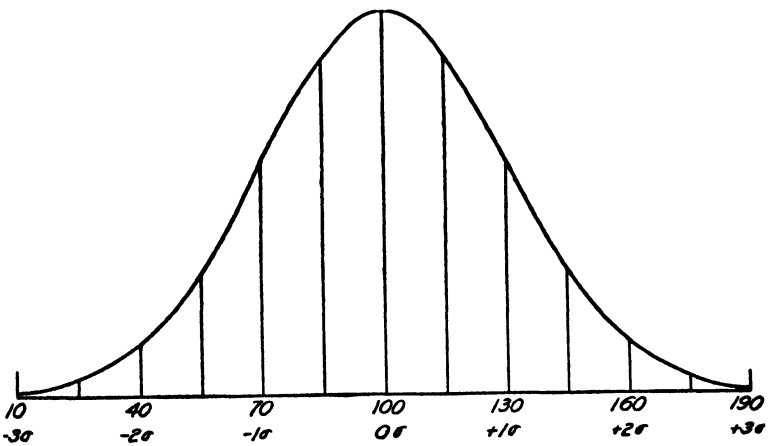


Figure 28. Frequency curve showing standard deviation as unit of measurement on the base line

be designated as $-1\ \sigma$; 130 can be designated $+1\ \sigma$; 115 is $+\ .5\ \sigma$; 40 is $-\ 2\ \sigma$; and so on.

As has previously been pointed out, we are usually interested in two fundamental facts about a frequency distribution — the central tendency, which is a point on the X-scale, and the variability, which is a distance on the X-scale. The central tendency shows the general location of the numbers on the X-scale, *i.e.*, it shows whether a list of salaries ranges around $100 or $200 per month. The variabil-

ity shows how far the salaries scatter or how closely they concentrate about the central tendency. If a single case is considered, such as the number 160 in the distribution of *Figure 28*, we should specify both the central tendency and the variability of the distribution, in order to have a fair idea as to where the number 160 stands in relation to its fellows. Both of these facts are combined, as regards a single number, 160, by designating it $+ 2\sigma$. The fact that this notation is positive shows that the number is above the average, and the fact that it is 2σ shows that it is far above the average of the distribution. The total range of a symmetrical distribution represents about six sigma, three positive and three negative. Theoretically, the distribution extends to infinity above and below the mean, but over 99% of the cases in a symmetrical surface are included between the limits $+ 3\sigma$ and $- 3\sigma$.

When two frequency curves are compared as to their variability by means of the standard deviation, a large standard deviation indicates a greater variability than a small standard deviation. The properties of the standard deviation will be discussed more in detail in connection with the normal frequency curve.

The standard deviation is used so extensively in statistical work that several methods have been devised for calculating it. We shall describe three of these methods:

1. Without class intervals
2. With class intervals and an arbitrary origin
3. In terms of the original numbers

The choice of the method of calculation depends on one's individual preferences, the available calculating machines, and the arrangement of the given data.

1. Calculation of standard deviation without class intervals. In *Table 9* is illustrated the simplest procedure for calculating the standard deviation without any short cuts. In the first column are numbered all the observations or scores. There are 29 cases in

#	X	d	d^2
1	5	0.7	0.49
2	7	1.3	1.69
3	3	2.7	7.29
4	6	0.3	0.09
5	7	1.3	1.69
6	4	1.7	2.89
7	1	4.7	22.09
8	5	0.7	0.49
9	6	0.3	0.09
10	4	1.7	2.89
11	5	0.7	0.49
12	7	1.3	1.69
13	2	3.7	13.69
14	6	0.3	0.09
15	6	0.3	0.09
16	3	2.7	7.29
17	10	4.3	18.49
18	5	0.7	0.49
19	4	1.7	2.89
20	5	0.7	0.49
21	7	1.3	1.69
22	12	6.3	39.69
23	8	2.3	5.29
24	4	1.7	2.89
25	7	1.3	1.69
26	6	0.3	0.09
27	5	0.7	0.49
28	8	2.3	5.29
29	8	2.3	5.29
	$\Sigma X = 166$		$\Sigma d^2 = 147.81$

$$\text{Mean, } m = \frac{\Sigma X}{n} = \frac{166}{29} = 5.72$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{147.81}{29}} = 2.26$$

Table 9. Calculation of standard deviation without class intervals

this illustration. In the second column are listed the numbers themselves in whatever order they may occur in the original data. The sum of the second column is ΣX, which is in this case 166. This sum, $\Sigma X = 166$, is divided by the number of cases, $n = 29$, which gives the mean, 5.7. In the third column are recorded the deviations from this mean, disregarding sign. In the fourth column are listed the squares of the individual deviations. The sum of this column is $\Sigma d^2 = 147.81$. The remaining substitutions in the formula for the standard deviation are shown in *Table 9*. This method of calculating the standard deviation becomes unwieldy with a large number of cases and with relatively large numbers.

2. Calculation of standard deviation with class intervals and arbitrary origin. In *Table 10* is illustrated another method of calculating the standard deviation. It is perhaps more generally used than any other method. In the first column are listed the class intervals, so defined as to avoid ambiguity in classification. In the second column we list the corresponding class frequencies. The sum of the f column is the number of cases, n, which in this illustration is 140.

The next step is to assume an arbitrary origin somewhere in the middle range of the distribution. That class interval is chosen as the arbitrary origin and labeled zero which is nearest the mean, as far as this can be guessed by inspection. The arbitrary origin can be located anywhere on the scale, even

outside the range, without in any way affecting the arithmetical accuracy of the calculation, but the numbers involved in the calculation are made the smallest possible by locating the arbitrary origin at or near the mean. This facilitates the arithmetical work. In the illustration of *Table 10* we have called the class-interval 90–99 zero and labeled it zero in the *d* column. The other spaces in the *d* column are labeled 1, 2, 3, 4, etc. in both directions from the zero class interval. The deviations are therefore calculated in terms of class intervals and not in terms of the original scale units. The class intervals higher

Class Intervals	*f*	*d*	*fd*	fd^2
40–49	1	− 5	− 5	25
50–59	5	− 4	− 20	80
60–69	12	− 3	− 36	108
70–79	21	− 2	− 42	84
80–89	23	− 1	− 23	23
90–99	23	0	− 126	
100–109	25	+ 1	+ 25	25
110–119	14	+ 2	+ 28	56
120–129	11	+ 3	+ 33	99
130–139	4	+ 4	+ 16	64
140–149	1	+ 5	+ 5	25
	140		+ 107	589

$n = 140$
$\Sigma fd + = 107$
$\Sigma fd - = 126$
$\Sigma fd = -19$
$\Sigma fd^2 = 589$

$$c = \frac{\Sigma fd}{n} = \frac{-19}{140} = -0.136$$

$c^2 = 0.018$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{n} - c^2} \text{ class intervals}$$

$$= \sqrt{\frac{589}{140} - 0.02} \text{ class intervals}$$

$= 2.04$ class intervals
$= 20.4$ scale units

Table 10. Calculation of standard deviation with class intervals and an assumed origin

than the zero interval are designated as positive deviations, and the intervals lower than the zero class interval are designated as negative deviations, as shown in the *d* column of *Table 10*.

The *fd* column contains the products of the *f* and *d* columns. The sum of the negative items in the *fd* column is Σfd_{neg}, which in the illustration is -126. The sum of the positive items in the *fd* column is Σfd_{pos}, which is 107 in the illustration. The difference between these two items is Σfd, which in the illustration is -19.

In the fd^2 column we have the products of the *d* and *fd* columns. The sum of this column is Σfd^2, which is 589 in the illustration.

The correction, *c*, is necessary on account of the use of an arbitrary or guessed origin during the calculations. It is determined and used as shown in *Table 10*.

The standard deviation in this illustration is 2.04 class intervals, but we must note that each class interval contains ten scale units, as is seen by the column of class intervals. Therefore, the standard deviation is 2.04 *class intervals* or 20.4 *scale units*.

3. Calculation of standard deviation in terms of the original numbers. In *Table 11* we have a method of calculating the standard deviation in terms of the original numbers, so as to avoid dealing with deviations and correction for the arbitrary origin. This method seems simple at first, but it is labor-saving only when the numbers dealt with are relatively small. Even in that case one must be careful to

carry the calculations to a sufficient number of significant digits to insure reasonable accuracy in obtaining a relatively small difference between two relatively large numbers under the radical of this formula.

#	X	X^2
1	5	25
2	7	49
3	3	9
4	6	36
5	7	49
6	4	16
7	1	1
8	5	25
9	6	36
10	4	16
11	5	25
12	7	49
13	2	4
14	6	36
15	6	36
16	3	9
17	10	100
18	5	25
19	4	16
20	5	25
21	7	49
22	12	144
23	8	64
24	4	16
25	7	49
26	6	36
27	5	25
28	8	64
29	8	64
	166	1098

$$\Sigma X = 166$$

$$n = 29$$

$$m = \frac{\Sigma X}{n} = \frac{166}{29} = 5.72$$

$$m^2 = 32.7$$

$$\Sigma X^2 = 1098$$

$$\sigma = \sqrt{\frac{\Sigma X^2}{n} - m^2}$$

$$= \sqrt{\frac{1098}{29} - 32.7}$$

$$= \sqrt{5.14}$$

$$\sigma = 2.26$$

Table 11. Calculation of standard deviation in terms of the original numbers

In the first column is shown the numerical order of the observations or scores. There are 29 cases in this illustration. In the second column are listed the scores themselves. Their sum is 166. In the

third column are listed the squares of the numbers. Their sum is 1098. The remaining calculations are shown in *Table 11*.

In general, the second method of calculating the standard deviation involves the least amount of labor.

The formula for the standard deviation used in the third method gives the same result as the formulæ used for the first two methods. The identity can be shown as follows:

The deviation, d, of any particular number, X, from the mean, m, of the series is

$$d = X - m$$

Squaring, we have

$$d^2 = X^2 - 2mX + m^2$$

Hence the summation,

$$\Sigma d^2 = \Sigma X^2 - 2m\Sigma X + nm^2$$

Dividing by n,

$$\frac{\Sigma d^2}{n} = \frac{\Sigma X^2}{n} - 2m\frac{\Sigma X}{n} + m^2$$

$$= \frac{\Sigma X^2}{n} - 2m^2 + m^2$$

$$= \frac{\Sigma X^2}{n} - m^2$$

Hence

$$\sigma = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{\Sigma X^2}{n} - m^2}$$

Problem 1. Calculate the standard deviation for the three curves in *Figure 25*.

Problem 2. A distribution has the following constants: $m = 76.4$; $\sigma = 14.36$. What is the value of X, which is indicated as -1.62σ?

Chapter Sixteen

Percentile Ranks

It is sometimes desirable to state the relative standing of an individual with reference to the other members of a group. This is usually done by giving him a rank. Thus, if there are seventeen persons in a group, the person who ranks highest is given a rank of seventeen and the person whose rank is lowest is given a rank of one.

If we are told that a person has a rank of 27, we do not know whether his rank is high, average, or low, unless we also know the number of individuals in the group. If a person ranks 27 in a group of 29, he ranks high relative to the group. If he ranks 27 in a group of 1,000, his rank is relatively low. When designating an individual by his absolute rank, one must also state the number of individuals in the group with which he is compared. The absolute rank of an individual member of a group is his numerical position when the ranks of all the members of the group have been arranged in order from the lowest to the highest. The highest numerical rank is given to the person who ranks highest. In this sense the designation of absolute ranks differs from the customary expression of calling the best person "first." In statistical language "one" means the

lowest numerical score, no matter what the meaning of the score may be.

In order to avoid the necessity of stating the number of cases involved in designating a person's relative position by rank, one may express the absolute rank in terms of *percentile rank*. In this case one states the rank that the person would have if there were one hundred members in the group. If there are fifty individuals in a group, the middle person would have an absolute rank of 25, but his percentile rank would be 50. A percentile rank is so calculated that it indicates the per cent of the group which ranks below the specified percentile. A person who has a percentile rank of 72 exceeds 72 per cent of the group and is exceeded by 28 per cent of the group.

The median score is always the 50th-percentile. The upper quartile is always the 75th-percentile, and the lower quartile is always the 25th-percentile.

When several individuals have the same score, a certain assumption is made concerning their ranks. Consider the ten individuals designated by the following scores:

Individual:	A	B	C	D	E	F	G	H	I	J
Score:	16	24	24	35	41	56	56	56	72	83

These scores are arranged from the lowest to the highest. Now if we designate these individuals and their scores by absolute ranks, we have the following arrangement:

Individual:	A	B	C	D	E	F	G	H	I	J
Score:	16	24	24	35	41	56	56	56	72	83
Absolute Rank:	1	2	3	4	5	6	7	8	9	10

But B and C have the same score, and consequently they should have the same rank in order to state fairly their relative position in the group with which they are compared. This is done by assigning to each of these two persons the same rank; namely, their median rank. The revised statement of absolute ranks is then as follows:

Individual:	A	B	C	D	E	F	G	H	I	J
Score:	16	24	24	35	41	56	56	56	72	83
Absolute Rank:	1	$2\frac{1}{2}$	$2\frac{1}{2}$	4	5	7	7	7	9	10

In this case B and C, who have the same score, 24, are given the same absolute rank, namely, $2\frac{1}{2}$; and the three individuals, F, G, and H, who have the same score, 56, are given the same absolute rank, namely, 7. The total number of ranks assigned is ten because there are ten individuals in the group. No individual is given the rank of 6 or 8, these being skipped in order to balance the ranks for the small groups of equal scores.

The same assumption is made concerning percentile ranks. Suppose that one hundred individuals are ranked from the lowest to the highest and that the twenty highest individuals all have the same score. It would be unfair to assign a percentile of 81 to one of these twenty persons and a percentile of 100 to another when both have the highest possible score. In cases of this sort all twenty are given the same percentile rank, namely, 90, which is the median percentile rank of the twenty individuals who have the same score at the top of the group. The same assumption is made for groups of identical scores in the

middle range of the distribution and at the lower end. As a consequence it is possible to assign percentile ranks in such a way that no individual gets a percentile of 100 even though several of them have the highest possible score.

In the accompanying table, *Table 12*, we have the calculation of the percentile ranks for the Liberal Arts freshmen at the University of Texas in an intelligence test. The first two columns indicate the class intervals, two columns being used to specify the lower and upper limits of each class interval. The third column gives the midpoint of the class interval. The fourth column gives the frequencies. The sum of the frequency column is of course the total number of cases, n, which in this illustration happens to be 860.

The next step is to label the three percentile columns, the lower percentile, the upper percentile, and the midpercentile, respectively. This is done because the percentile rank which belongs to the bottom of a class interval is not the same as that which belongs to the upper edge of that class interval. The midpercentile between these two is assigned as the percentile rank for the whole class interval. The failure to take this bit of logic into consideration accounts for many inconsistencies in the calculation of percentile ranks.

The first step in the actual calculation is to determine the rate, which is the reciprocal of the total number of cases. In the illustration of the accompanying example, the total number of cases is 860, which is n.

The rate is $\frac{1}{n}$ or $\frac{1}{860}$, which is .001162. This can be determined either by performing the indicated division or directly from Barlow's Tables or any other similar tables.

Score			Frequency	Percentiles		
From	To	Mid		Lower	Upper	Mid
1	*2*	*3*	*4*	*5*	*6*	*7*
0	9	5	0	....	....	....
10	19	15	0	....	....	....
20	29	25	6	.000	.006	.003
30	39	35	19	.006	.029	.017
40	49	45	56	.029	.094	.061
50	59	55	94	.094	.203	.148
60	69	65	154	.203	.382	.292
70	79	75	161	.382	.569	.475
80	89	85	143	.569	.735	.652
90	99	95	95	.735	.845	.790
100	109	105	65	.845	.921	.883
110	119	115	33	.921	.959	.940
120	129	125	16	.959	.978	.968
130	139	135	8	.978	.987	.982
140	149	145	7	.987	.995	.991
150	159	155	3	.995	1.000	.997
160	169	165	0	....	....	....
	.....	*Total*	860	....	....	....

Rate = $1/n$ = 1/860 =0.001162

Table 12. Calculation of percentile ranks

Every class interval represents a percentile range. In the fifth column we have the lower limit and in the sixth column the upper limit of the percentile range for each class interval. The upper limit of the percentile range for any given class interval is identical with the lower limit of the percentile range of the

next higher class interval. This may be seen by inspecting the tabulated calculations. The percentile rank which is finally given to all the individuals in any class interval is the midpoint of the percentile range for that class interval. This is tabulated in the last column.

After the rate has been determined, one registers this number in a calculating machine. In the illustration here given the rate would first be added six times (corresponding to the first frequency of 6); and the machine would then show .006, which is the upper percentile rank for the class interval 20–29. The lower percentile rank for the lowest class interval is of course zero, and the upper percentile rank for the highest class interval is necessarily 1.00. Add the rate nineteen times more in the machine and it will then show .029. This is in reality nothing but cumulative addition. The upper limit, .006, for the class interval 20–29 is identical with the lower limit, .006, for the class interval 30–39. Record these percentiles and add the rate cumulatively fifty-six times. This will show .094 as the upper percentile for the next class interval.

When the whole column has been completed, one will have added the rate n times (.001162 × 860), which will give 1.00 for the upper limit of the highest class interval. The calculation of percentile ranks is greatly facilitated by using a calculating machine, such as the Burroughs adding machine with the attachment by which the setting may be shifted to the right and to the left.

Percentile ranks are calculated by adding the rate cumulatively, as indicated by the frequency column. The sum thus obtained immediately before adding the frequency for any given class interval is the lower percentile rank for that class interval. The sum obtained after the class frequency times the rate has been added is the upper percentile rank for that class interval.

With the help of an adding machine, the calculation of percentile ranks, such as those of the illustration, requires only three or four minutes. Any clerk can be taught to do this work; and the method here described affords an automatic check on the accuracy, because the last sum must necessarily be unity.

The percentile curve is a curve which shows the relation between the given variable, such as scores, and percentile ranks. By means of a percentile curve, one may ascertain at a glance the percentile rank corresponding to any given score. The rank is of course with reference to the group whose distribution is represented by the curve. *The percentile curve is in reality a cumulative frequency curve in which the frequency ordinates have been stated as fractions of n.*

In *Figure 29* we have the percentile curve for the frequency table and percentile calculations in *Table 12*. The base line represents scores in the test, and the ordinates represent percentile ranks. By consulting the table we find, for example, that the class interval 80–89 has a percentile rank of .65. This agrees with the curve, which has an elevation of .65 for the midpoint of the class interval 80–89.

Note that the class interval 80–89 has a corresponding percentile range from .569 to .735. This range is indicated on the graph by a short vertical line extending from the level .569 to the level .735. We can therefore read directly from the chart the percentile range for each class interval. Note that this vertical

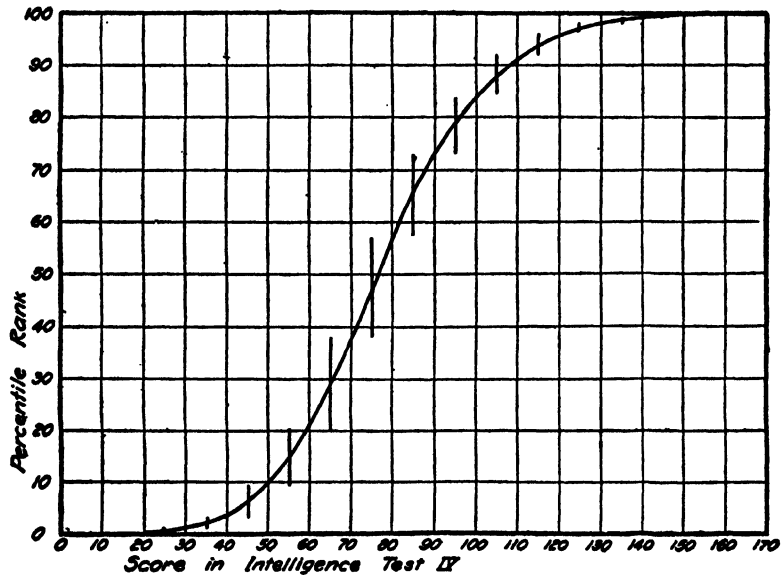


Figure 29. Percentile curve corresponding to Table 12

line is drawn not at either end of the class interval but directly above its midpoint. The curve is plotted by drawing these vertical lines as indicated by columns 5 and 6 in the table. The curve is drawn through the midpoints of these vertical lines. In practice one may therefore eliminate the calculation of the mid-percentiles, since these can be plotted by inspection

with sufficient accuracy for all practical purposes. When one becomes accustomed to this work, it is possible to eliminate the third, sixth, and seventh columns in the table and to determine these facts graphically.

In practice it is found that with a small number of cases the curve would not be smooth if it were drawn strictly through the midpoints of the vertical lines. For this reason one sometimes generalizes the curve by smoothing it so as to pass as nearly as possible through the midpoints of the successive vertical lines. This gives a *smoothed percentile curve.* The reasoning by which a percentile polygon is smoothed into a percentile curve is similar to that of the smoothing of a frequency polygon into a smoothed frequency curve. However, when a percentile curve has been smoothed, it is only fair to the reader of one's report that the vertical lines be inserted, because they represent the actual observations. The reader can ascertain at a glance to what extent the curve has been smoothed by noting the deviation of the curve from the midpoints of the vertical lines. It should also be observed that the percentile polygon is more readily seen to be continuous than the corresponding frequency polygon. Even when the frequency polygon is quite irregular, the percentile polygon or curve which is plotted on the same drawing and to the same scale will be relatively continuous or smooth.

We may summarize the properties of the percentile curve as follows:

1. The percentile curve is asymptotic to the base line or zero and also to the percentile rank of 100. The percentile curve theoretically never reaches zero nor 100. Therefore it is theoretically impossible for any individual to have a percentile rank of zero or a percentile rank of 100. This fact is probably not generally recognized.

2. The median score can be read directly from a percentile curve because it is the score which corresponds to the 50th-percentile. In the illustration we find that the 50th-percentile gives a score of 77, which is the median of the group. This saves considerable labor in calculating the balanced median.

3. The upper quartile can be determined in a similar manner because the upper quartile point is the 75th-percentile. In the illustration the 75th-percentile corresponds to a score of 92. The lower quartile point corresponds to the 25th-percentile, which in the illustration is 63. When the percentile curve has been plotted, it is therefore not necessary to calculate the quartile points because these can be read directly from the graph. From these facts one may of course also determine graphically the quartile range or the semi-interquartile range, as desired.

4. The mode is defined as the point on the base line of a frequency curve directly under the highest ordinate of the curve. The mode is therefore that score which occurs most frequently in the distribution. This can be determined approximately from

the percentile curve by selecting the point on the base line directly under the steepest part of the percentile curve. Since the percentile curve is in reality a cumulative frequency curve, it should be clear that the steepest part of the percentile curve is directly over that score which occurs with the highest frequency. While this graphical determination is only approximate, it may be sufficiently accurate for practical purposes and it saves considerable labor in computation.

5. When two distributions are to be compared, considerable information may be obtained by simply inspecting the percentile curves for the two distributions. Thus, for example, the ranges of the distribution may be compared by inspecting the ranges of the two percentile curves. The variabilities of the two distributions may be compared by inspecting the relative slopes of the two percentile curves. The steeper the general slope of the percentile curve, the smaller is the variability of the corresponding distribution. The two medians may also be compared graphically.

When a percentile curve is to be constructed, it is not necessary to perform all the calculations indicated in the accompanying table of data. That table was prepared so that it would be complete and so that it would show all the steps that are implied in the calculations.

In practice it is necessary to calculate only the figures in columns 1, 2, 4, and 5. But when the calculations are made with an abbreviated form, it

is absolutely necessary to keep in mind the fact that the percentiles recorded in the fifth column are not the percentiles that are to be assigned to the adjacent class intervals. The percentiles recorded in the fifth column indicate the lower percentile of the percentile range for each class interval. In order to locate the upper percentile for any class interval, it is necessary to consult the percentile recorded in column 5 for the next higher class interval. This becomes easy with practice because the figures are then immediately transferred to a chart, which facilitates the interpretation.

Graphical calculation of percentile ranks.[1] It is possible to calculate percentile ranks without any numerical calculation whatever. A graphical method is illustrated in *Figure 30*. Let us suppose that we have a frequency table with the scores arranged in class intervals and the corresponding class frequencies. Let there be 156 cases in the distribution, arranged in class intervals of ten along the range from zero to 170. Plot the accumulated frequencies in the form of vertical lines over the midpoints of their respective class intervals. The lengths of these lines would be equal to the actual frequencies according to the frequency scale at the right end of the diagram. Draw a smooth curve, *A*, through the midpoints of these vertical lines. This is the accumulated frequency curve. In the diagram the lines representing the class frequencies have been omitted

[1] The section on graphical calculation of percentile ranks may be optional in the lesson plan.

in order to avoid confusion. The accumulated frequency curve *A* can be used to determine the absolute rank of any member of the group. What we want, however, is a similar curve which gives the

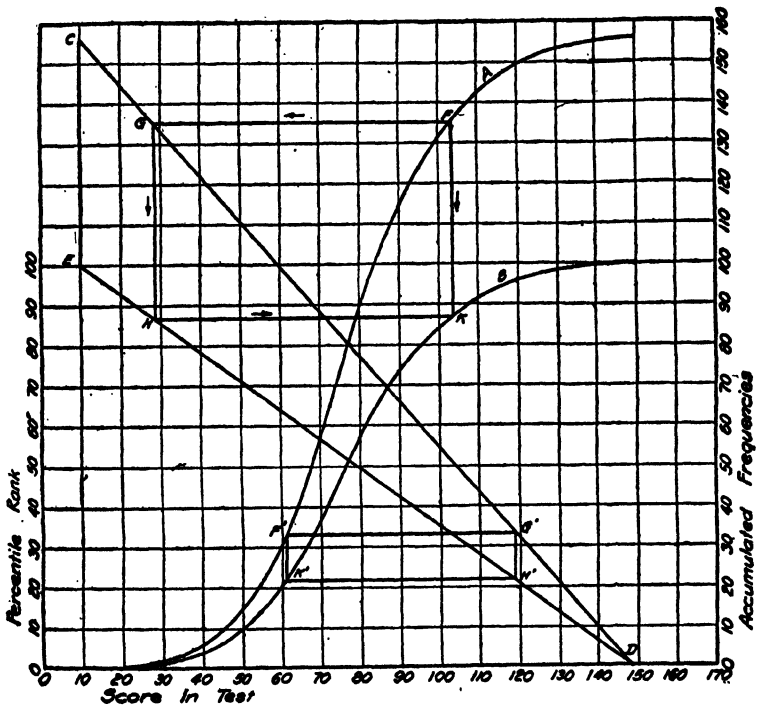


Figure 30. A graphical method of calculating percentile ranks

percentile rank instead of the absolute rank. Draw on the left edge of the diagram any suitable scale running from zero to 100 to represent percentile ranks.

The graphical construction is made as follows. Locate any point, *C*, at random on the elevation of

the top of the curve *A*. Locate the point *E* directly under the point *C* and on the elevation of 100 on the percentile scale at the left of the diagram. From these two points, *C* and *E*, draw straight lines to any point, *D*, on the base line of the diagram.

Assume any point, *F*, on the curve *A*. Draw a horizontal line from this point to its intersection with the upper diagonal line, at *G*. Drop a vertical line from *G* to its intersection with the lower diagonal line, at *H*. Draw a horizontal line from *H* to a point, *K*, directly under the starting point, *F*. This point *K* is the point on the percentile curve corresponding to the point *F* on the curve of absolute ranks. Proceed similarly with a number of other points on the curve *A*. In the lower part of the diagram the same procedure has been indicated for another point, *F'*, on the curve *A*.

Draw a smooth curve through the points so located. This will be the desired percentile curve, *B*. The procedure is very easily carried out on a drawing board, in fact, much more easily than the verbal description would indicate. In plotting the original curve, *A*, of accumulated frequencies, one should be careful to draw it through the midpoints of the vertical lines which represent the class frequencies.

An automatic check on the plotting of this curve is obtained in the fact that the total elevation of the vertical lines for curve *A* should equal the total number of cases in the distribution, which in the illustration is 156.

Where suitable calculating machines are available, this graphical procedure is not recommended. It is then best to deal with the numerical frequencies and the numerical forms of the percentile ranks as indicated in the accompanying frequency table.

It is essential that those who use percentile ranks in the tabulation of educational and psychological measurements should understand the fundamental properties of the percentile curve. If one becomes accustomed to the correct calculation of the percentile ranks, he will not be inclined to tolerate rough-and-ready guesses for the percentile ranks that should be assigned to given scores. The most frequent error in the assignment of percentile ranks to given scores is in the form of a constant error. This is caused by the fact that the percentile is determined either for the upper edge of a class interval or for the lower edge. This percentile is used for the whole class interval instead of the one which corresponds closely to the midpoint of the interval. The error is avoided by following the procedure illustrated in *Table 12* with percentile calculations.

Problem 1. Assume that you have before you a table with two columns. In the first column are the examination marks for a group of students. In the second column are the percentile ranks. On the assumption that the distribution of marks is normal, how would you estimate approximately the standard deviation of the distribution of marks?

Problem 2. Prepare a table of percentile ranks for each of *Tables 3, 4, 9,* and *14.*

Problem 3. Draw the percentile curves corresponding to *Tables 3, 4, 9,* and *14.*

Problem 4. Assume that you have before you the names of one thousand students with the percentile rank of each student in an examination. Draw a freehand sketch to show the general appearance of the frequency curve of these percentile ranks. Draw another sketch on the same base line to show the general appearance of the distribution of five hundred percentile ranks drawn at random from the list of one thousand names.

Problem 5. Assume that there are ten students in a class and that they have different marks in an examination. What would be their percentile ranks? Remember that no one can have a percentile rank of 0 or 100.

Problem 6. Make a freehand sketch of two percentile curves on the same chart to represent two frequency distributions, *A* and *B*, which have the same mean and the same number of cases. The distribution *A* has a wider scatter than *B*. Label the two curves.

Problem 7. Make a freehand sketch of two percentile curves on the same chart to represent two frequency distributions, *A* and *B*, which have the same mean and standard deviation. The distribution *A* has a larger number of cases than *B*. Label the two curves.

Problem 8. Make a freehand sketch of two percentile curves on the same chart to represent two frequency distributions, *A* and *B*, which have the same standard deviation and the same number of cases. The distribution *A* has a higher mean than *B*. Label the two curves.

Problem 9. The table on the following page represents the distribution of stature of 750 freshmen at Ohio State University in 1913. Calculate the percentile ranks corresponding to the specified heights. State the percentile rank of a student whose stature is 5 feet 10 inches. Explain just what is meant by his percentile rank in this case.

Height in Inches	j
61	2
62	10
63	11
64	38
65	57
66	93
67	106
68	126
69	109
70	87
71	75
72	23
73	9
74	4
	750

Chapter Seventeen

The Binomial Expansion

The frequency distributions of mental abilities and of anthropometric measurements conform more or less closely to the so-called probability curve. The statistical methods of treating distributions which approximate the symmetrical form of the probability curve necessitate some consideration of the statistical properties of the curve.

Permutations. The different ways in which a number of things can be arranged in a series are called *permutations*. Let us consider the two things *a* and *b*. These can be arranged in two ways, *ab* and *ba*. These two arrangements are called permutations.

Let us consider the three things *a*, *b*, and *c*. If we group these three things, taking only two at a time, we have six possible permutations, namely,

ab	*ba*	*ca*
ac	*bc*	*cb*

Each of three things is here combined successively with one of the other things. If we group the three things, taking three at a time, we have six permutations, namely,

abc	*bac*	*cab*
acb	*bca*	*cba*

Let us consider four things, a, b, c, d. If we group these four things, taking two at a time, we have twelve permutations, namely,

ab	*ba*	*ca*	*da*
ac	*bc*	*cb*	*db*
ad	*bd*	*cd*	*dc*

Here, as before, each one of the things is combined with each of the other things. If we group the four things, taking three at a time, we have twenty-four permutations, namely,

abc	*abd*	*acd*	*bcd*
acb	*adb*	*adc*	*bdc*
bac	*bad*	*cad*	*cbd*
bca	*bda*	*cda*	*cdb*
cab	*dab*	*dac*	*dbc*
cba	*dba*	*dca*	*dcb*

Similarly, if we group the four things, taking four at a time, we have twenty-four permutations.

The number of possible permutations of n things, taken r at a time, can be predicted by the following formula:

$$P_{nr} = n\,(n-1)\,(n-2)\,(n-3)\cdots(n-r+1)$$

The preceding examples may be verified with this formula. Thus the number of permutations of four things, taken three at a time, is $4\,(4-1)\,(4-2) = 24$.

Combinations. Combinations differ from permutations in that one combination such as *abc* may be stated in the form of several permutations by simply rearranging the order in which the things occur in

the series, such as *abc*, *acb*, *bac*, *bca*, *cab*, *cba*. All of these are considered to be one combination, but they represent six permutations. In any given situation the number of permutations is always greater than the number of possible combinations.

Let us consider the two things *a* and *b*. The number of combinations of these two things, taken two at a time, is only one, namely, *ab*. If we consider the three things *a*, *b*, *c*, taking two at a time, we have three possible combinations, namely, *ab*, *ac*, *bc*. If we consider the three things, taking three at a time, we have only one possible combination, namely, *abc*.

Let us consider the four things *a*, *b*, *c*, *d*. If we group these four things, taking two at a time, we have six possible combinations, namely, *ab*, *ac*, *ad*, *bc*, *bd*, *cd*. If we group the four things, taking three at a time, we have four possible combinations, namely, *abc*, *abd*, *acd*, *bcd*. If we group the four things, taking four at a time, we have only one possible combination.

The number of combinations of n things, taking r at a time, can be predicted by the following formula:

$$C_{nr} = \frac{n(n-1)(n-2)(n-3)\cdots(n-r+1)}{r!},$$

in which $r!$ represents factorial r or $(1 \times 2 \times 3 \times 4 \times 5 \cdots r)$. Thus the number of combinations of four things, taking three at a time, is

$$C_{4\cdot3} = \frac{4(4-1)(4-2)}{1\cdot2\cdot3} = \frac{4\cdot3\cdot2}{1\cdot2\cdot3} = 4.$$

It may assist the student in his understanding of combinations and permutations to consider the following example. Suppose that you have bought two articles. When you have bought them, your total purchase will constitute one *combination* of two things. But you may have bought them in different orders. The order of the separate purchases may be *ab* or *ba*. These orders are called *permutations*. Thus we have two permutations of the two things, but only one combination of them.

Now consider yourself buying two things, with three things from which to choose. How many combinations are possible? The number of combinations is the number of purchases that you can make with two things in each purchase. If the three things are *a*, *b*, *c*, and if you are limited to the purchase of any two of them, you will have three possible purchases, namely, *ab*, *ac*, *bc*. In statistical language, we should say that there are three combinations of three things, taken two at a time. The two things in each purchase may be bought in different order. In statistical language, we should say that there are six permutations of three things, taken two at a time.

Probability of single event. Suppose that fifty white balls and fifty black balls are placed in a bag. If we draw one ball at random from the bag, the probability is $\frac{1}{2}$ that it will be a white ball and the probability is $\frac{1}{2}$ that it will be a black ball. The sum of the two probabilities is equal to unity or certainty. We can state this relation as follows:

$$P_B + P_W = 1.$$

If the bag contains 16 white balls, 20 black balls, 24 green balls, and 28 red balls, there is a total of 88 balls in the bag. Now if we draw 1 ball at random from the bag, the probability is $\frac{16}{88}$, or .182, that it will be a white ball. The probability is $\frac{24}{88}$, or .273, that it will be a green ball. The probabilities for the remaining colors appearing in a single draw are indicated in the following equation:

$$P_W + P_B + P_G + P_R = 1,$$

or

$$.182 + .227 + .273 + .318 = 1.$$

The *probability* is a ratio between the number of possibilities in favor of any specified color and the total number of possibilities involved. Since there are only four colors represented by the balls in the bag, it follows that the sum of the probabilities for the four colors must equal unity or certainty.

If we are interested in drawing a white ball from the bag and prefer to avoid drawing one of the colored balls, then the probability of drawing a white ball is called the probability of success, which in this case is .182. The probability of failure is the probability of drawing a colored ball, which is the sum of the remaining probabilities, namely, .818. The probability of success would in this case be roughly as one to five. If one were betting, one would come out even in the long run by betting one to five on white.

The probability of success is usually designated by the letter P, and the probability of failure is designated by the letter Q, so that $P + Q = 1$.

Probability of compound event. Suppose that the bag contains fifty white balls and fifty black balls. If we draw two balls at random from the bag, we may have the following permutations:

1. White, white
2. White, black
3. Black, white
4. Black, black

These probabilities represent three combinations, namely,

1. Two white balls
2. Two black balls
3. One white and one black ball

The *probability of compound events* refers to such problems as the probability of drawing two white balls in succession, one white ball and one black ball in two draws, four white balls in five draws, etc.

What is the probability of drawing two white balls in succession? The probability of drawing a white ball in the first draw is .5 and the probability of drawing a white ball in the second draw is also .5, if we assume that the first ball is returned to the bag before the second draw is made. Our problem now is to determine the probability of drawing two white balls in succession. We draw a white ball in the first draw in only one-half of the cases. Of these successful first draws, one-half will be eliminated by the fact that the second draw is a black ball in one-half of the cases. Hence we should expect to draw two white balls in succession in one-fourth of the

cases. This can be generalized by saying that the probability of drawing two white balls in succession is the product of the probabilities of the two draws, namely, $.5 \times .5 = .25$.

If the bag contains 20 white balls and 80 black balls, the probability of drawing a white ball is .2. The probability of drawing two white balls in succession is $.2 \times .2 = .04$. The probability of drawing three white balls in succession is $(.2)^3 = .008$; in other words, this would happen only about eight times in every one thousand complete draws. The probability of drawing three black balls in succession is $(.8)^3 = .512$.

The binomial expansion. Our object now is to show the significance of the terms of the binomial expansion in connection with the probability for compound events. The binomial expansion is as follows:

$$(P + Q)^n = P^n + \frac{n}{1} P^{n-1} Q + \frac{n(n-1)}{1 \cdot 2} P^{n-2} Q^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} P^{n-3} Q^3 + \frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4} P^{n-4} Q^4 + \cdots$$

After the first two or three terms have been written, the continuity of the series becomes apparent, so that the remaining terms can be written. The notation is as follows:

P = the probability of success
Q = the probability of failure
n = the number of events

If we apply this equation to the probability for compound events, such as tossing four coins, the significance of the several terms is as follows: the first term represents the probability that all four coins will fall heads; the second term represents the probability that three of the four coins will fall heads; the third term represents the probability that two of the four coins will fall heads; the fourth term represents the probability that one of the four coins will fall heads; the fifth term represents the probability that all four coins will fall tails. The sum of these five probabilities is equal to certainty, which is designated unity.

Let us analyze the first five terms of the expansion with special reference to the tossing of four coins. The five terms represent respectively the probabilities of getting four heads, three heads, two heads, one head, and no heads. The sum of these five terms is unity when four events are considered.

The first term. If a coin is tossed once, the probability of having it fall heads is $\frac{1}{2}$ and the probability of having it fall tails is $\frac{1}{2}$. The probability of getting four coins in succession to fall heads is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = (\frac{1}{2})^4 = \frac{1}{16}$. This principle has been established in the preceding section on the probability for compound events. Since we are considering heads as success and tails as failure, the notation for this probability becomes P^n, in which $P = \frac{1}{2}$ and $n = 4$, which gives $\frac{1}{16}$. Therefore, in the long run, we should expect to find that one out of every sixteen throws of four coins will give us four heads. This is the first term of the binomial expansion.

The second term. The probability of all heads in the first three tosses is $P^3 = (\frac{1}{2})^3 = \frac{1}{8}$ and the probability of tails in the fourth toss is $Q = \frac{1}{2}$. Therefore the probability of getting heads in the first three tosses and tails in the fourth toss is $P^3Q = \frac{1}{8} \times \frac{1}{2} = \frac{1}{16}$. But what we really want to know is the probability of getting heads in any three of four tosses. This implies that the one coin which may fall tails can be the first coin, the second coin, the third coin, or the fourth coin. Therefore, we should expect that " heads in any three of four tosses " will occur four times as frequently as " heads in the first three tosses and tails in the fourth toss." The possibilities of heads in any three of four tosses may be expressed as follows:

H H H T
H H T H
H T H H
T H H H

We have already seen that the probability of three heads in the first three tosses and tails in the fourth is $\frac{1}{16}$. The probability of heads in the first two tosses is $(\frac{1}{2})^2 = \frac{1}{4}$, and the probability of tails in the third toss is $\frac{1}{2}$. Therefore the probability of getting heads in the first two tosses and tails in the third toss is $P^2Q = \frac{1}{8}$. The probability of getting heads in the last toss is $\frac{1}{2}$. Thus the probability of getting heads in the first two tosses, tails in the third toss, and heads in the fourth toss is

$$P^2QP = \frac{1}{4} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}.$$

Similarly it can be shown that the probability of getting heads in the first toss, tails in the second, and heads in the last two is $\frac{1}{16}$, and also that the probability of getting tails in the first toss and heads in the last three tosses is $\frac{1}{16}$. The sum of these four probabilities, $\frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{4}$, is the probability of getting heads in any three of four tosses. This probability of heads in any three of four tosses is given by the second term of the binomial expansion, which is

$$nP^{n-1}Q = 4(\tfrac{1}{2})^3(\tfrac{1}{2}) = 4 \cdot \tfrac{1}{16} = \tfrac{4}{16} = \tfrac{1}{4}.$$

The third term. The probability that the first two tosses will be heads is $P^2 = (\frac{1}{2})^2 = \frac{1}{4}$. The probability that the last two tosses will be tails is $Q^2 = (\frac{1}{2})^2 = \frac{1}{4}$. The probability that the first two tosses will be heads and the last two tosses tails is the product of these two probabilities, namely, $P^{n-2}Q^2$. This accounts for a part of the third term of the binomial expansion. But what we really want is the probability of heads in any two of four tosses. The probability of heads in any two of four tosses will be greater than $\frac{1}{16}$ because the two heads may appear in any one of the following combinations: 1–2, 1–3, 1–4, 2–3, 2–4, 3–4. These numbers indicate tosses. The probability of heads in any two of four tosses is therefore $\frac{6}{16}$. The factor 6 is determined by the coefficient $\frac{n(n-1)}{1 \times 2}$, in which $n = 4$. This factor is the number of combinations in which four things can be arranged, taking two at a time. It is the formula

described in the preceding section on combinations. The third term in the binomial expansion represents the probability of heads in any two of four tosses, when $n = 4$, $P = \frac{1}{2}$, and $Q = \frac{1}{2}$.

The fourth term. The probability that the first toss will be heads is $P^{n-3} = \frac{1}{2}$. The probability that the last three tosses will be tails is $Q^3 = (\frac{1}{2})^3 = \frac{1}{8}$. Therefore the probability that the first toss will be heads and the last three tosses tails is

$$P^{n-3}Q^3 = (\tfrac{1}{2})(\tfrac{1}{2})^3 = (\tfrac{1}{2})^4 = \tfrac{1}{16}.$$

As before, however, what we really want is the probability of a head in any one of four tosses. This implies that the three tails may occur in any combination in the four tosses. Now, the number of combinations of four things, taking three at a time, is

$$\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} = \frac{4 \cdot 3 \cdot 2}{1 \cdot 2 \cdot 3} = 4.$$

Therefore the occurrence of a head in any one of four tosses is four times as frequent as the occurrence of a head on only the first of the four tosses. The probability of a head in any one of four tosses is represented by the fourth term in the binomial expansion. This term is

$$\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} \cdot P^{n-3}Q^3 = 4 \cdot (\tfrac{1}{2}) \cdot (\tfrac{1}{2})^3 = \tfrac{4}{16} = \tfrac{1}{4}.$$

The fifth term. The probability of having four tails in four tosses is $Q^4 = (\frac{1}{2})^4 = \frac{1}{16}$. The coefficient for the fifth term in the expansion is

$$\frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 2 \cdot 3 \cdot 4} = 1.$$

This coefficient represents the number of combinations of four things, taking four at a time. It turns out to be 1, and this agrees with common sense because there can obviously be only one combination of four things, taking all four of them at a time. The probability of four tails in four tosses is therefore $\frac{1}{16}$.

It is important to notice that the five terms add up to unity. The numerical form of the equation is as follows:

$$(P + Q)^n = \tfrac{1}{16} + \tfrac{4}{16} + \tfrac{6}{16} + \tfrac{4}{16} + \tfrac{1}{16} = 1.$$

Probability of 4 3 2 1 0 heads in four tosses.

Six tosses. As an illustration of the application of the binomial expansion to a longer series of tosses we have in the following outline the result for six tosses. The method is exactly the same. Two additional terms of the expansion are necessary and these are written by extending the continuity of the series.

All the possibilities for six tosses are represented in the following list:

1. Six heads, no tails
2. Five heads, one tail
3. Four heads, two tails
4. Three heads, three tails
5. Two heads, four tails
6. One head, five tails
7. No heads, six tails

One sees here that, with six tosses in the series, seven terms are required to represent the possible combinations of heads and tails. These seven terms

of the binomial expansion are as follows (compare with expansion in a preceding paragraph for five terms):

$$(P+Q)^n = P^n + \frac{n}{1}P^{n-1}Q + \frac{n(n-1)}{1\cdot 2}P^{n-2}Q^2 + \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}P^{n-3}Q^3 + \frac{n(n-1)(n-2)(n-3)}{1\cdot 2\cdot 3\cdot 4}P^{n-4}Q^4 + \frac{n(n-1)(n-2)(n-3)(n-4)}{1\cdot 2\cdot 3\cdot 4\cdot 5}P^{n-5}Q^5 + \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6}P^{n-6}Q^6 + \cdots$$

The numerical evaluation of these terms is indicated in *Table 13*. Note that the sum of all the probabilities is equal to unity.

The student will probably find *Table 13* very useful in studying the binomial expansion. Notice, for example, that the probability of having all heads in six tosses is the same as the probability of having all tails in six tosses. Similarly, the probability of having four heads in six tosses is the same as the probability of having four tails (two heads) in six tosses. These relations should be mastered not only as far as the algebraic notation is concerned but also as to their reasonableness.

In *Figure 31* we have represented the probabilities for these six events in graphical form. Notice that we have a curve of familiar shape, high in the central region and low at the extremes. The shaded

Order of Terms	*Algebraic Notation*			*Value of Coefficient*	*Factor in P*	*Factor in Q*	*Probability*	*Interpretation*
	Coefficient	*P*	*Q*					*Probability of Getting*
1		P^n		1	$\times \frac{1}{64}$		$= \frac{1}{64} = 0.016$	Six heads in six tosses
2	$\frac{n}{1}$	$\times P^{n-1}$	$\times Q$	6	$\times \frac{1}{32}$	$\times \frac{1}{2} = \frac{6}{64} = 0.094$		Five heads in six tosses
3	$\frac{n(n-1)}{1 \cdot 2}$	$\times P^{n-2}$	$\times Q^2$	15	$\times \frac{1}{16}$	$\times \frac{1}{4} = \frac{15}{64} = 0.234$		Four heads in six tosses
4	$\frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}$	$\times P^{n-3}$	$\times Q^3$	20	$\times \frac{1}{8}$	$\times \frac{1}{8} = \frac{20}{64} = 0.312$		Three heads in six tosses
5	$\frac{n(n-1)(n-2)(n-3)}{1 \cdot 2 \cdot 3 \cdot 4}$	$\times P^{n-4}$	$\times Q^4$	15	$\times \frac{1}{4}$	$\times \frac{1}{16} = \frac{15}{64} = 0.234$		Two heads in six tosses
6	$\frac{n(n-1)(n-2)(n-3)(n-4)}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}$	$\times P^{n-5}$	$\times Q^5$	6	$\times \frac{1}{2}$	$\times \frac{1}{32} = \frac{6}{64} = 0.094$		One head in six tosses
7	$\frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}$	$\times P^{n-6}$	$\times Q^6$	1	$\times 1$	$\times \frac{1}{64} = \frac{1}{64} = 0.016$		No heads in six tosses

Table 13. Interpretation of the binomial expansion

rectangle in the figure represents $\frac{1}{64}$, as may be seen from the x- and y-axes. If we should obtain the area of the surface under the curve by a planimeter, it would be very nearly equal to unity.

We have now plotted the theoretical probability curve. The curves that we have plotted before were

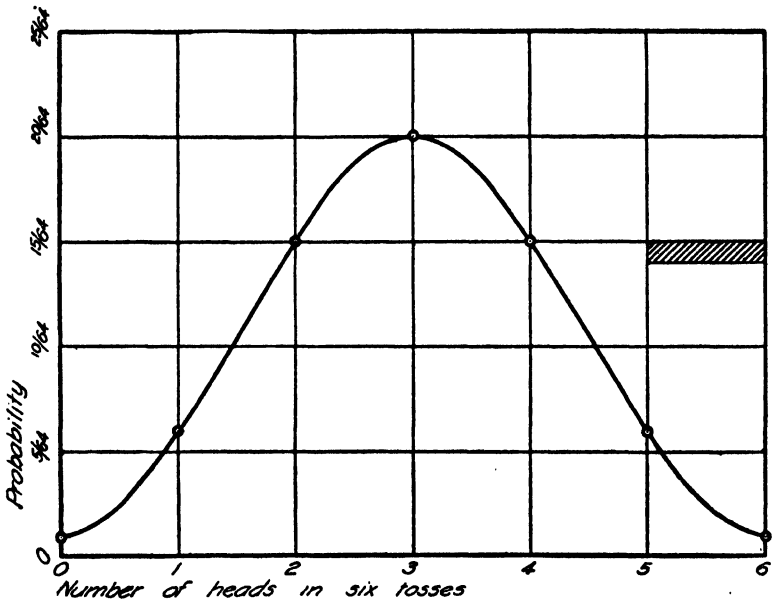


Figure 31. Probabilities for six tosses

obtained from empirical data which contain chance deviations from the theoretically expected curve. It will be possible now to compare a curve obtained by actual data with the curve that could have been expected by chance. Such comparisons are sometimes of fundamental significance in statistical work.

Refer to the preceding table for six tosses. Note, for example, that the probability of getting two heads in six tosses is $\frac{15}{64}$ or .234. If six coins are tossed in one thousand trials, we should expect approximately 234 of the trials to give two heads.

Problem 1. Calculate by means of the binomial expansion the probabilities of getting one, two, three, etc. heads in eight tosses of a coin. This will necessitate the use of nine terms of the expansion. Plot a theoretical probability curve for your calculations. If a coin were tossed on 648 trials with eight tosses on every trial, in how many of these trials would you expect to have five heads in eight tosses?

Problem 2. If some one offers to bet ten dollars to one of yours that you cannot get five heads or more in six tosses, and if he is willing to keep on betting this way, how much would you expect to gain or lose after one hundred such bets? (See table in text for six tosses.)

Problem 3. Show what is wrong with the following reasoning. "The probability of drawing a red card from a deck of cards is $\frac{1}{2}$, and the probability of drawing a black card is also $\frac{1}{2}$ on the assumption that the joker has been removed. Therefore the probability of drawing five red cards in ten successive draws is also $\frac{1}{2}$."

What is the probability of drawing five red cards in ten draws? Assume that the card which is drawn is returned to the deck before the next card is drawn.

Problem 4. Assume that a bag contains 200 balls in the following proportions: 40 red, 50 green, 50 blue, 40 black, 20 white. Designate the red, green, and blue balls as colored. Calculate the probability of drawing one, two, three, etc. colored balls in six draws. Arrange your calculations as shown in the calculations for six tosses in the text. Note that your values for P and Q in this problem will not be identical as they are in the text problem. Plot the probabilities. How does this curve vary from the usual probability curve? Show that the shape

of this curve could have been foreseen as reasonable by logical considerations before plotting the curve.

Problem 5. Assume that you have ten cards before you, and that five of them are red, five of them are black. The cards are shuffled and you are to draw five cards from the pile of ten cards. Assume also that you do not return each card to the pile after it has been drawn. Calculate the probability of drawing five red cards from the pile. Note that this probability is less than $(\frac{1}{2})^5$. Why is this reasonable?

Chapter Eighteen

The Probability Curve

The curve represented by the successive terms of the binomial expansion is called the *probability curve* or the *normal curve.* It can be stated in the form of the equation

$$y = y_0 e^{\frac{-x^2}{2\sigma^2}}, \qquad (1)$$

in which

x = the deviation from the mean in terms of the standard deviation,

y = the ordinate and represents expected frequency,

σ = the standard deviation of the distribution in terms of class intervals,

e = the constant 2.718, known as the Napierian base,

y_0 = the ordinate at the mean.

This equation enables one to determine the ordinate of the probability curve at any deviation from the mean.

The ordinate at the mean is expressed by the relation

$$y_0 = \frac{n}{\sigma\sqrt{2\pi}} = \frac{n}{2.5066\sigma}, \qquad (2)$$

in which

n = the number of cases in the distribution,

σ = the standard deviation of the distribution in terms of class intervals,

π = the constant 3.1416.

By means of this equation one can determine the expected frequency at the mean in terms of the number of cases and the standard deviation, expressed in terms of the class intervals.

The more complete form of the equation for the probability curve is therefore

$$y = \frac{n}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}, \qquad (3)$$

which enables one to superimpose a normal probability curve on a given frequency distribution to determine how closely the experimental data conform to the normal probability curve.

In order to do this one must first calculate the expected frequency at the mean by equation (2). The standard deviation and the number of cases are determined in the usual way. When the expected frequency at the mean has been determined by equation (2), one can determine from *Table 19*, in the appendix, the ordinates of the remainder of the curve. These ordinates are given in the table as fractions of the ordinate at the mean. Thus, for example, we find from the table that the ordinate at $1\,\sigma$ is .606. If the ordinate at the mean is 100, the ordinate at $1\,\sigma$ would be 60.6 and similarly for other points on the curve. The curve is symmetrical about the mean and therefore the ordinates for positive values of σ are identical with the ordinates for the corresponding negative values of σ. Thus the ordinate of the probability curve at $-.74\,\sigma$ is .760 of the ordinate

at the mean. If the ordinate at the mean were 100, the ordinate at $-.74\sigma$ would be 76.

Let us suppose that a frequency distribution has been plotted as in *Figure 32*. We desire to superimpose the probability curve on this distribution in order to see how far the actual distribution deviates from the normal curve. The small circles in the

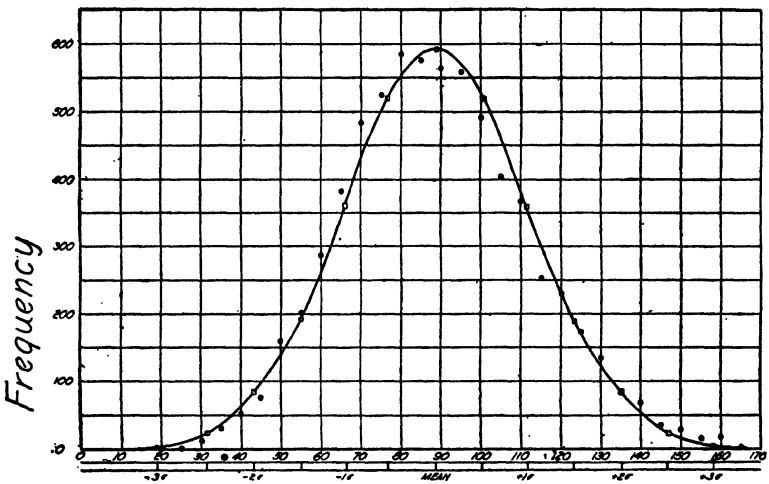


Figure 32. Normal curve superimposed on a frequency polygon

figure represent the actual frequencies and they correspond to the frequencies recorded in *Table 14*. The successive operations in superimposing the probability curve on the given distribution are as follows:

1. Calculate the standard deviation as shown in *Table 14*. Note that since σ is there calculated in terms of column *d*, the σ will be expressed in terms of class intervals. Hence $\sigma = 4.61$ class intervals or

23.05 test score units, there being five score units in each class interval.

2. Calculate the true mean of the distribution as shown in *Table 14*. The true mean is 88.9 score units. The sigma scale has its origin at this point, as shown in *Figure 32*.

X	f	d	fd	fd^2
20	2	13	26	338
25	2	12	24	288
30	14	11	154	1,694
35	30	10	300	3,000
40	53	9	477	4,293
45	76	8	608	4,864
50	162	7	1,134	7,938
55	202	6	1,212	7,272
60	287	5	1,435	7,175
65	378	4	1,512	6,048
70	483	3	1,449	4,347
75	524	2	1,048	2,096
80	580	1	580	580
85	575	0	9,959	
90	564	1	564	564
95	561	2	1,122	2,244
100	492	3	1,476	4,428
105	405	4	1,620	6,480
110	368	5	1,840	9,200
115	256	6	1,536	9,216
120	233	7	1,631	11,417
125	173	8	1,384	11,072
130	137	9	1,233	11,097
135	82	10	820	8,200
140	69	11	759	8,349
145	35	12	420	5,040
150	31	13	403	5,239
155	14	14	196	2,744
160	16	15	240	3,600
165	2	16	32	512
	6806		15,276	

$n = 6{,}806$

$\Sigma fd + = 15{,}276$

$\Sigma fd - = 9{,}959$

$\Sigma fd = 5{,}317$

$\Sigma fd^2 = 149{,}335$

$$c = \frac{\Sigma fd}{n} = .781 = \text{correction for } \sigma$$

$c^2 = .61$

Calculation of σ

$$\sigma = \sqrt{\frac{\Sigma fd^2}{n} - c^2} = \text{class intervals}$$

$$= \sqrt{\frac{149335}{6806} - .61}$$

$\sigma = 4.61$

Calculation of mean

$$c = \frac{I \cdot \Sigma fd}{n} = \frac{5 \times 5317}{6806} = +3.9$$

c = correction for m.

$m = m_a + c = 85 + 3.9 = 88.9$

Table 14. Calculation of the mean and the standard deviation for a frequency table. The distribution represents intelligence test scores for 6806 engineering students

3. Calculate the mean ordinate of the probability curve which has a standard deviation of 4.61 and a total of 6806 cases.

$$y_0 = \frac{n}{2.5066\,\sigma} = \frac{6806}{2.5066 \times 4.61} = 590$$

The expected frequency at the mean is 590 when the distribution is plotted in class intervals of five.

4. Determine the points on the x-scale which correspond to $+ .5\,\sigma$, $+ 1\,\sigma$, $+ 1.5\,\sigma$, etc. This has been done in the following tabulation. The true mean is 88.90 and, since $\sigma = 23.05$, $\frac{1}{2}\,\sigma = 11.52$.

Mean	=	88.90	Mean =	88.90	
		+ 11.52		− 11.52	
+ .5 σ	=	100.42		77.38	= − .5 σ
		+ 11.52		− 11.52	
+ 1. σ	=	111.94		65.86	= − 1 σ
		+ 11.52		− 11.52	
+ 1.5 σ	=	123.46		54.34	= − 1.5 σ
		+ 11.52		− 11.52	
+ 2 σ	=	134.98		42.82	= − 2 σ
		+ 11.52		− 11.52	
+ 2.5 σ	=	146.50		31.30	= − 2.5 σ
		+ 11.52		− 11.52	
+ 3 σ	=	158.02		19.78	= − 3 σ

Consult *Figure 32* and verify the location of these points.

5. Calculate the expected ordinates at these points. This is done as shown in the tabulation on the following page.

σ	$\frac{y}{y_0}$	y
0	1.000	590
$+ .5\,\sigma$	.882	520
$+ 1\,\sigma$	.606	358
$+ 1.5\,\sigma$	.324	191
$+ 2\,\sigma$	.135	80
$+ 2.5\,\sigma$	.044	26
$+ 3\,\sigma$	.011	6

The fractions in the column headed $\frac{y}{y_0}$ are obtained from *Table 19*. The expected frequencies in the y column are fractional parts of the mean ordinate 590. Note by *Figure 32* that these frequencies, y, are identical for positive and negative values of σ, the probability curve being symmetrical.

Verify several of the ordinates on the probability curve in *Figure 32*. We can now see to what extent the actual frequency distribution of the 6806 test scores deviate from the probability curve. Note that the test score distribution is slightly positively skewed. The deviation from the normal surface is relatively slight but noticeable.

Table 19 can be used not only to determine the ordinate for given deviations from the mean but also for determining the deviation from the mean at which a given frequency may be expected when the frequency at the mean is known. Thus if we want to determine the deviation from the mean at which the expected frequency is one-half of the frequency at the mean, we find .50 in *Table 19*, which shows that a frequency one-half of the frequency at the mean can be expected at $1.18\,\sigma$ and at $-1.18\,\sigma$.

Problem 1. Construct three probability curves on the same sheet according to the following specifications. Indicate an ordinate at the midpoint of each class interval.

Curve	σ	Mean	n	Class Interval
A	15	50	400	10
B	15	50	800	10
C	15	50	1200	10

Problem 2. Construct three probability curves on the same sheet according to the following specifications. Indicate the ordinates by small circles at each half-sigma.

Curve	σ	Mean	n
A	10	60	400
B	20	60	800
C	30	60	1200

Problem 3. Construct three probability curves on the same sheet according to the following specifications. Indicate ordinates by small circles at each half-sigma.

Curve	σ	Mean	n
A	10	60	1000
B	20	60	1000
C	30	60	1000

Problem 4. A given distribution has the following constants: $m = 56.3$; $n = 4320$; $\sigma = 11.62$. Assume that the distribution is normal. 1. What would you expect the mean ordinate to be? 2. At what X-value do you expect the ordinate to be one-third of the mean ordinate? 3. What ordinate do you expect at the X-value 41.5?

Problem 5. Plot the data (the stature of college freshmen) of Chapter Sixteen, *Problem 9*, on cross section paper and indicate the ordinates of the data in the form of small circles. Superimpose a normal probability curve so as to show graphically the similarity between the distribution of stature and the probability curve. Indicate on the base line the mean, $.5\,\sigma$, $+1\,\sigma$, $+1.5\,\sigma$, $+2\,\sigma$, $+2.5\,\sigma$ for both positive and negative deviations as shown in *Figure 32*. Compare the actual distribution with the superimposed probability curve.

Chapter Nineteen

The Area of the Frequency Surface

We have previously noted that the area of the frequency surface is proportional to the number of cases represented. If the area under the probability curve is assumed to be unity, it is possible to state the ratio of the total number of cases between any two specified points on the X-scale to the total number of cases in the distribution. In *Table 20* these ratios are tabulated.

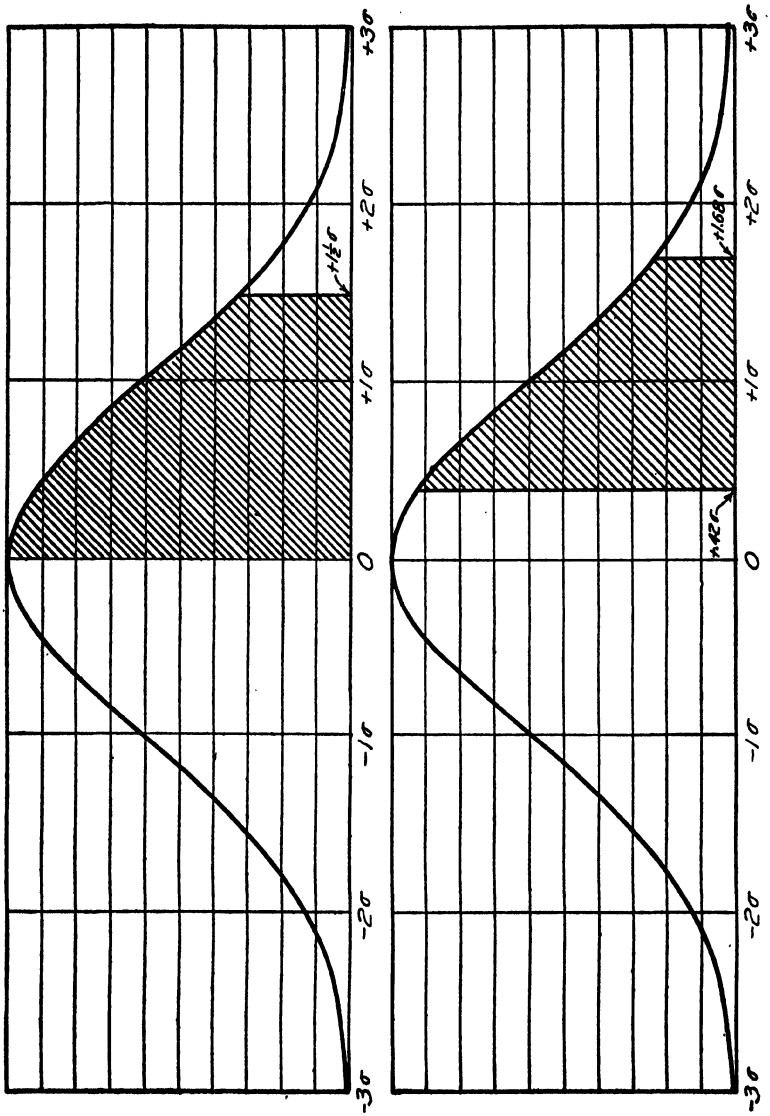
Suppose that we want to know what fractional part of the whole surface lies between the mean and $+ 1.5\,\sigma$. This part of the probability surface is represented by the cross-hatched area of the first diagram in *Figure 33*. By reference to *Table 20* we find the fraction .4332, which indicates that 43% of the total number of cases in a normal surface are located between the mean and $+ 1.5\,\sigma$. If the total number of cases, n, is 500, the number of cases between the mean and $+ 1.5\,\sigma$ would be $500 \times .43$ or 215 cases. The same procedure is followed for other similar problems.

Now suppose that we desire to know the number of cases between $+ .42\,\sigma$ and $+ 1.68\,\sigma$ in a normal distribution with 460 cases. This part of the probability surface is also represented by a cross-hatched dia-

gram in *Figure 33*. By reference to *Table 20* we find that 16.28% of the cases lie between the mean and $+ .42\sigma$ in a normal surface. Similarly we find from the same table that 45.35% of the total distribution is to be found between the mean and $+ 1.68\sigma$. The difference between these two percentages, 45.35 − 16.28, is 29.07. Hence 29.07% of the total number of cases lies between $+ .42\sigma$ and $+ 1.68\sigma$. Since n is 460, the answer to our problem is $460 \times 29.07 = 134$, which is the number of cases to be expected between $+ .42\sigma$ and 1.68σ.

Suppose that we want to know the number of cases between the mean and $- .65\sigma$ in a normal distribution of 280 cases. See *Figure 33*. We find that 24.22% of the total number of cases lie between the mean and $- .65\sigma$. Note that, since the curve is symmetrical about the mean, there will be the same number of cases between the mean and $- .65\sigma$ as between the mean and $+ .65\sigma$. Since n is 280, the answer to our problem is $280 \times 24.22 = 67.8$.

Still another combination is the problem to determine the number of cases to be expected between, say, $- .48\sigma$ and $+ 2.12\sigma$. See *Figure 33*. To solve this we divide the surface into two parts, the division being made at the mean. We find that 18.44% of the whole surface is in the negative part of the shaded area and 48.30% of the surface is in the positive part of the shaded area. Adding these two percentages, we have $18.44 + 48.30 = 66.74\%$, as the percentage of the whole surface between $- .48\sigma$ and 2.12σ.


+3σ
+2σ
+1σ
0
-1σ
-2σ
-3σ
+1.68σ
+3σ
+2σ
+1σ
0
-1σ
-2σ
-3σ
+1.68σ
MEAN

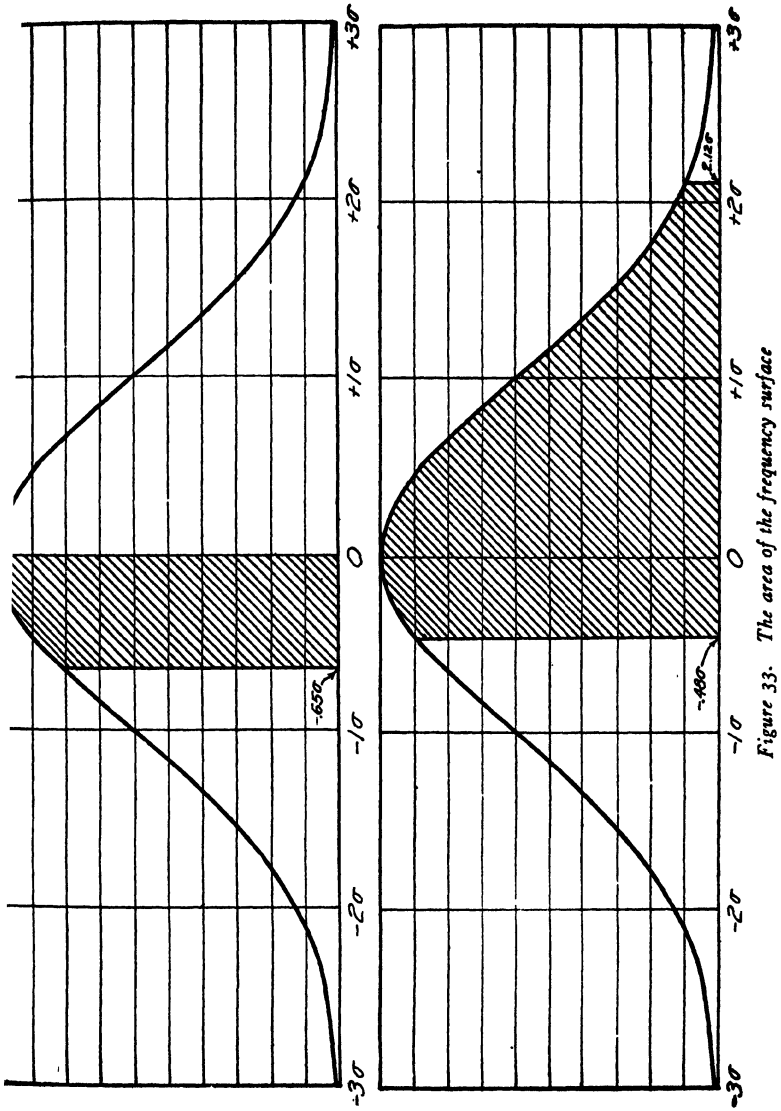


Figure 33. The area of the frequency surface

Problem 1. Plot a normal curve on a base line extending from -3σ to $+3\sigma$. Divide this base line into five equal parts. Determine the percentage of cases over each of the equal parts of the base line.

Problem 2. Plot a normal curve and mark the base line into five parts so chosen that each of the five parts will have an equal number of cases.

Problem 3. Determine approximately what percentage of the normal surface extends above $+3\sigma$ or below -3σ.

Problem 4. A normal surface has the following constants: $m = 76.34$; $\sigma = 2.43$ class intervals; $I = 10$; $n = 785$. How many cases do you expect to find between the X-values of 62 and 94? (I is the number of X-scale units in each class interval.) Express 62 and 94 in terms of σ deviations.

Chapter Twenty

Transmutation of Measures

The distinction between ranks and standings is a fundamental one in statistical work. If the individuals in a group are to be studied as to the distribution of stature, for example, we may designate the relative position of any member of the group with reference to the rest of the group by stating his absolute or percentile rank or by the actual variable in terms of inches, in terms of deviation in inches from the mean, or in terms of the standard deviation. In the first case we are concerned with *ranks* and in the second case we are concerned with *standings*.

The distinction between ranks and standings can perhaps be made clear by an example. If all the one hundred individuals in a group are arranged according to stature from the shortest to the tallest and assigned absolute ranks, the difference between any two adjacent individuals will be the same *in terms of ranks*. But in terms of stature the difference between any pair of adjacent individuals will not be the same. The difference in stature between two adjacent individuals will be relatively great at the extremes and relatively small in the middle range of the distribution. The actual difference in stature in inches represented by the percentile range 40 to 60 is not nearly so great as the actual difference in stature

represented by the percentile range 80 to 100. This is because the frequencies in the middle class intervals are usually greater than the frequencies at the extremes.

When we translate the data from one of these forms to the other, the procedure is known as transmutation of measures. If the data are given in the form of percentile ranks, we are unable to say anything regarding the form of the distribution. It may be normal or skewed. The ranks tell us nothing about this. When it is desirable to express a series of ranks in the form of standings, one generally assumes that the distribution is normal.

Suppose that two judges have given estimates of the same group of individuals and that it is desired to combine the two judgments for each person into a single combined judgment. If the two judges had in mind the same scale, the combination could be made directly, but usually one judge is more lenient than the other, or he uses more differentiating steps in his estimates. Let both judges use the same symbols, *A*, *B*, *C*, *D*, for their estimates, with *A* as the highest estimate, and *D* as the lowest estimate. If one judge gives twice as many grades of *A* as the other judge, it would not be fair to consider the grades of the two judges as equivalent. If one individual is rated *C* by one judge and *B* by the more lenient judge, our problem is to assign for statistical treatment some single score to this individual based on the two separate judgments.

We assume that the abilities judged are distributed

in the form of a normal probability surface. For the purpose of illustration let there be one hundred individuals in the group. Let the number of grades assigned by each judge be as follows:

	Judge 1	Judge 2
Grade A	40	20
Grade B	30	30
Grade C	20	30
Grade D	10	20
	$n = 100$	$n = 100$

It is apparent that Judge 1 grades more leniently than Judge 2.

The transmutation of measures is a very simple operation. First make a rough pencil sketch similar

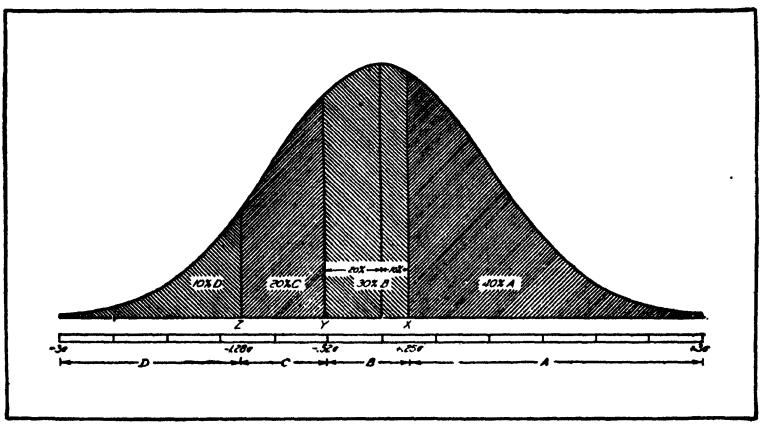


Figure 34. Transmutation of measures

to *Figure 34*, to represent the distribution of grades assigned by Judge 1. The sketch should contain the outline of the normal frequency surface and vertical lines to indicate the letter grade groups. In this

particular problem there are four such groups. The *A* grades are given to forty per cent of the group and hence will be indicated on the sketch by a section almost as large as half of the total frequency surface.

Letter Grades	*Per Cent of Whole Group*	*Per Cent Limits*	*Average*	*Corresponding Sigma Standing*
A	40	+ 50 + 10	+ 30%	+ .84 σ
B	30	+ 10 − 20	− 5%	− .13 σ
C	20	− 20 − 40	− 30%	− .84 σ
D	10 n = 100	− 40 − 50	− 45%	− 1.65 σ

Table 15. Transmutation of measures

The *B* grades are given to the next thirty per cent and that section is smaller and contains the mean or median. The other sections are indicated on the sketch in a similar manner. From the sketch it can now be seen that the middle person in the *A* group is removed thirty per cent from the median of the whole group. The sigma value for the point in the surface which is removed thirty per cent of the whole group from the median is .84 σ, as determined from *Table 20*. In *Table 15* we have the calculation indicated for Judge 1. A similar table is made for each judge. The sigma equivalents

for the letter grades given by the two judges are tabulated below.

SIGMA EQUIVALENTS FOR LETTER GRADES

	Judge 1	Judge 2
Grade A	.84 σ	1.28 σ
Grade B	− .13 σ	.39 σ
Grade C	− .84 σ	− .39 σ
Grade D	− 1.65 σ	− 1.28 σ

If we desire to combine the estimates of the two judges, we can now do so by assigning the numerical equivalents of the letter grades and averaging them. Thus if one of the individuals in the group was given a grade of *A* by Judge 1 and a grade of *C* by Judge 2, his combined standing would be

$$\frac{+ .84 - .39}{2} = + .22$$

This is the best possible way of combining the separate judgments. It assumes that the distribution of the abilities estimated is normal, and it also assumes that we place equal confidence in the judgments of each judge. These assumptions may or may not be true, but they serve as the connecting link by which the two separate judgments are brought together into one combined judgment for statistical treatment.

This procedure enables one not only to combine disparate judgments in a reasonable way but it also enables one to state in objective form the standards used by the judges. The procedure can be applied to such varied data as the scholarship grades of teachers, judgments on rating scales, and estimates of abilities to be used as criteria for determining the

predictive value of mental tests. There is an almost universal tendency to overestimate abilities no matter who the judge may be. This is appropriate in its place but it is troublesome in statistical analyses. All of the members of a group cannot possibly be above the average of the group. A recent newspaper account of illiteracy in the draft army made the alarming statement that one-half of a regiment was below the average of the regiment!

Problem 1. The following is the distribution of estimates of three judges:

	Judge 1	Judge 2	Judge 3
Grade A	20	12	35
Grade B	22	22	22
Grade C	19	32	20
Grade D	24	27	17
Grade E	21	13	12
	106	106	106

Determine the sigma standing to be assigned to each grade by each judge for the purpose of combining the estimates. Prepare a table of such equivalent standings. What would be the sigma standing of an individual who is rated *B* by judge 1, *A* by judge 2, and *D* by judge 3? What is the maximum sigma standing possible with this transmutation, and what is the minimum possible standing?

Problem 2. Prepare a table with two columns to show the relation between percentile ranks and sigma standing for a normal distribution. Make the entries in the percentile column 0, 5, 10, 15, 20, etc., and determine the corresponding standings from the tables in the appendix.

Prepare a similar table with the same headings. Make the entries in the sigma columns -3σ, -2.5σ, -2.0σ, -1.5σ, etc., and determine the corresponding percentile ranks from the tables.

Show in some graphical way the distinction between ranks and standings.

Chapter Twenty-one

The Probable Error

Reliability of statistical measures. Everybody is figuring probable errors. Statistical jobs in education are being justified as scientific, dignified, and trustworthy by the fact that probable errors have been figured. It is essential to ascertain what the probable error means, and what it does not mean, what it does show and what it does not show, the kind of reliability that it does give as well as the kind of assurance that it does not give at all. By noting the assumptions that are basic for the probable error, the constant can be intelligently used, and the mistake will not be made of assuming that just because the probable error is small the statistical values are therefore trustworthy.

Suppose that you want to know the average age of all sixth-grade children in the city of Chicago. The ideal and complete way to answer such a question would be to list the actual age of every sixth-grade child in that city. In most circumstances it is very difficult to obtain complete data for answering statistical questions. One has recourse to the data for a smaller but supposedly representative group. In the case of the illustration you would probably ascertain the ages of perhaps one thousand sixth-grade children and determine their average age.

You would use that figure as representative of the average for the whole city. The smaller group that you actually study, instead of the whole group that you would use in the ideal situation, is called a *sample.*

If the sample were one thousand sixth-grade children, their average age would be thought of as a fairly reliable figure because one thousand is a large number. If the sampling were one hundred children instead of one thousand, the average age would be thought of as less reliable. If the sample were reduced to ten children, their average age would be a relatively unreliable measure of the age of sixth-grade children in the whole city. Finally, if the sample were reduced to three or four children, their average age would be so unreliable that we should place no confidence in it whatever, as far as representing the city is concerned. It is clear that as we increase the number of cases for which an average, or arithmetic mean, is calculated, the reliability, or the confidence, that we attach to the figure is considerably enhanced. This is nothing but common sense. Stated in more technical language, we should say that the reliability of an average is a function of the number of cases in the sample.

If we have ascertained the ages of one hundred children and calculated their average, and if we are using this average to prove something, it is of course important to be able to say just how reliable that average is. Stated in more statistical form, the problem is to determine how much the average age

might be expected to fluctuate if we should repeat the tabulation for another group of one hundred children, and for still another group, and so on. It is obvious that the average ages of the successive groups of children, with one hundred in each group, would not be absolutely the same. There would be minor fluctuations by mere chance even if the other factors of selection were kept uniform.

The minor fluctuations in the average age of successive groups of one hundred sixth-grade children would be relatively small. The fluctuations in the average age of successive groups would be larger if the groups were made smaller. Finally, if the groups were reduced in size until there were only one child for each group, instead of one hundred, the fluctuations in the average age of successive groups would be as large as the fluctuations in the ages of individual sixth-grade children. The purpose of the probable error, as applied to the arithmetic mean, is to indicate the relative magnitude of the fluctuations that are to be expected if successive groups, or samples, are obtained for calculating the mean.

Different sources of unreliability. There are several kinds of errors to be considered in statistical work in education and in the social sciences. One kind of error in the arithmetic average of a lot of numbers is that which is brought about by mere chance. These errors are relatively slight when the groups are large. If the average age for one hundred children were found to be 11.86 years, it is probable that the next similar group of children would have an

average age more or less closely approximating this number, although it would probably not agree to the second or third decimal place.

There is another kind of error which causes much more serious disturbance in statistical work, namely, that which is brought about by unintentionally loading the sample so that it is not representative of what we are really trying to measure. If children of the sixth grade are studied in some particular school, or part of a city, it would usually be found that a similar group of children from some other part of the city, representing a different social class, would show markedly different constants as to age, height, weight, intelligence, and other measures. An error of sampling whereby the group that is studied does not represent that which it is supposed to represent in the statistical study, causes errors far more serious than those which are brought about by chance fluctuations. It should be borne in mind that the probable error gives assurance regarding the relative magnitude of the expected fluctuations which are due to mere chance, but the probable error tells us absolutely nothing regarding the other and more important factors in selection of the groups to be studied by which they are not representative of the purposes of the statistical inquiry. It is essential to recognize that the probable error takes into account the fluctuations which are, in most practical studies, the smallest and least important of the fluctuations which disturb the confidence that one attaches to statistical routine. For example, assume that sixth-grade chil-

dren are classified by name, alphabetically, and that the investigator selects at random a continuous stretch of one hundred names from the alphabetical list. This would seem to be a random selection of the group to be studied, because the first letter of the child's name does not seem to be related positively with any of the traits that might be inquired about in a statistical study. Such a procedure would, however, have the possibility of giving an unintentional loading of the sample with some nationality in the city population, that, in turn, might conceivably introduce traits which are represented more heavily in the sample than they should be in order to represent the city as a whole. Such errors of sampling, and hundreds of other factors which bear on the legitimacy of the selection of a representative group, are not even touched by the probable errors of the resulting figures. Throughout the study of the probable error, and the other statistical constants that measure the reliability or unreliability of statistical measures, one should bear in mind that they give us some indication of the expected degree of fluctuation in successive samples selected in the same way, and that this indication refers to the least important of the various causes that disturb the validity of statistical findings, namely, the factor of chance fluctuation.

An experiment with the probable error. Assume that one thousand numbers have been written on cards, one for each card. Let the total range of the numbers be from zero to 24 and assume that the

arithmetic mean is at 12 with a standard deviation of 4. Let these cards be thrown into a pile and assume that a sample of twenty numbers be drawn from the pile. We should expect the average of the twenty numbers to be close to 12 because that is the known average of the whole pile, but it is improbable that a sample of twenty of the cards would give a mean exactly twelve. It would fluctuate by chance so that the average of the first sample of twenty numbers might be 11.5 or 13.2, or any number in the general vicinity of 12.

The purpose of the probable error is to indicate the extent to which the average of *additional* samples of twenty numbers might vary from the average that we did obtain for the only sample that we did take. This expected chance variation to which the average is subject is numerically shown by the probable error which, for the arithmetic mean, is given by the formula

$$\textit{Probable error of arithmetic mean} = \frac{.67\,\sigma}{\sqrt{n}},$$

in which σ is the standard deviation of the distribution of the sample, and n is the number of cases in the sample, which is twenty.

In an experiment actually carried out as described above the average of the first sample of twenty numbers was 12.4 and the standard deviation of the first sample was 3.21, which gives, for the probable error, the value of .48. The mean of the sample, with its probable error, is usually written as follows:

Mean = 12.40 ± .48. It has various interpretations that we shall discuss and examine separately.

It will be noticed that the number of cases, *n*, occurs in the denominator of the formula. This agrees with our previous statement that the expected chance fluctuations of an arithmetic mean are reduced by increasing the number of cases in the sample. The standard deviation is in the numerator. If the numbers in the pile vary from zero to one million, we should of course expect greater fluctuation in the average of successive samples of twenty cards than if the numbers in the pile range only from zero to ten. This fact is consistent with the formula in that the standard deviation occurs in the numerator. If the variability of the numbers in the sample were increased, we should expect a greater chance fluctuation of the mean.

The probable error or expected chance fluctuation of the mean varies directly as the standard deviation of the sample and inversely with the number of cases.

In *Figure 35* we have a frequency distribution for one thousand numbers with an arithmetic mean at 12 and a standard deviation of 4. The total range is from zero to 24. Samples of twenty numbers were drawn from the one thousand numbers and their means calculated. These means were plotted as small dots in the lower part of the diagram. Fifty such samples, each with twenty numbers, were drawn from the large pile of one thousand numbers. After each sample was drawn, the twenty cards were replaced in the large pile so that each sample was

always drawn from the full pile of one thousand cards. The fifty dots show that the means of the samples of twenty numbers do not scatter as much as the individual numbers. All the averages range between 10 and 15, whereas the original numbers range from zero to 24. In other words, if there are a

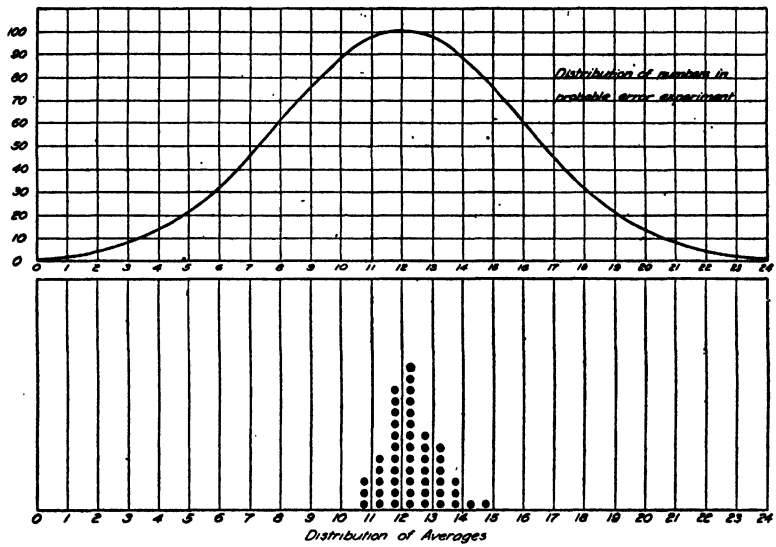


Figure 35. A probable-error experiment

few cards with numbers at the lower extreme of the range, such as 0, 1, 2, and a few numbers at the upper extreme end of the range, such as 22, 23, 24, we should not expect any sample of twenty numbers that we draw by chance from the whole pile to give us an average which is at either extreme. The averages do not vary as much as the individual numbers themselves; and, of course, the larger the group of num-

bers that we draw in the successive trials, the smaller will be the variation from one average to the next.

Samplings of Twenty Numbers	*Mean of the Sampling*	*σ of the Sampling*	*P. E. of Sampling*	*Is true mean, 12, inside P. E. limits?*	*Samplings of Twenty Numbers*	*Mean of the Sampling*	*σ of the Sampling*	*P. E. of Sampling*	*Is true mean, 12, inside P. E. limits?*
1	12.40	3.21	.48	Yes	26	12.60	4.40	.66	Yes
2	13.15	4.26	.64	No	27	10.50	3.49	.53	No
3	14.05	3.97	.60	No	28	12.05	3.40	.51	Yes
4	11.95	3.00	.45	Yes	29	11.55	3.81	.58	Yes
5	12.40	3.62	.55	Yes	30	12.25	4.06	.61	Yes
6	12.65	3.16	.48	No	31	12.25	3.74	.56	Yes
7	12.35	3.20	.48	Yes	32	11.05	2.91	.44	No
8	13.40	2.74	.41	No	33	12.55	3.41	.51	No
9	11.40	3.29	.50	No	34	11.35	3.80	.57	No
10	12.50	2.92	.44	No	35	11.65	3.23	.49	Yes
11	11.65	3.65	.55	Yes	36	12.40	3.06	.46	Yes
12	11.60	3.71	.56	Yes	37	13.60	3.04	.46	No
13	11.05	4.93	.74	No	38	11.45	4.48	.68	Yes
14	11.70	4.16	.63	Yes	39	12.05	5.37	.81	Yes
15	13.65	4.20	.63	No	40	11.50	3.19	.48	No
16	13.95	3.73	.56	No	41	12.80	6.44	.97	Yes
17	13.15	4.48	.68	No	42	10.55	3.38	.51	No
18	11.85	2.85	.43	Yes	43	13.00	3.66	.55	No
19	11.60	4.14	.63	Yes	44	12.55	4.21	.63	Yes
20	13.05	3.71	.56	No	45	12.15	4.04	.61	Yes
21	13.35	4.28	.65	No	46	12.15	4.76	.72	Yes
22	12.95	3.77	.57	No	47	12.05	4.15	.62	Yes
23	11.60	2.44	.37	No	48	11.95	3.67	.55	Yes
24	14.75	4.32	.65	No	49	12.45	4.13	.62	Yes
25	10.70	3.86	.58	No	50	12.40	3.90	.59	Yes

Table 16. An experimental study of the probable error

In *Table 16* we have the tabulation of records from the actual experiment with one thousand numbers from which fifty different samples were drawn at random with twenty numbers in each sample. In

the first column is listed the order of the successive samples as they were drawn. In the second column is listed the arithmetic mean of each sample. The extent of variation of the average for groups of this size and for the range of this experiment can be inferred from the second column.

The third column gives the standard deviation of each sample of twenty numbers. These standard deviations hover about 4, more or less, which is also the standard deviation of the large pile of numbers from which the samples were drawn. This verifies empirically that, if a sample be drawn, the standard deviation of the sample is comparable with the standard deviation of the universe which the samples represent. This becomes reasonable if we consider the ordinates of the frequency curve of *Figure 35*. The ordinate at 12 is 100, which means that there are 100 twelves in the large pile from which the samples are drawn. The ordinate at 8 is 61, which means that there are 61 eights in the large pile. Now, if a single card be drawn from the pile, it is clear that the chance of drawing a twelve is $\frac{100}{1000}$ or $\frac{1}{10}$, since there are one thousand cards in the pile. The probability of drawing an 8 is $\frac{61}{1000}$. In the long run we shall have, as the average shape of distribution curve for the samples of twenty cards, the same form as the distribution curve of the large pile, with the exception that since the number of cases is 20 in the sample, instead of 1000, we shall have a lower surface. The area of the surface representing the average sample will be $\frac{20}{1000}$ of the large surface which

represents the whole pile. The probability of drawing a twelve or an eight for the sample is exactly the same as the probability of drawing a twelve or an eight in constructing the large surface. Hence the variability of the sample will in general be the same, or comparable with, the variability of the large surface. The ordinates will all be proportionally lower to account for the fact that fewer numbers are being drawn in the sample.

In the fourth column we have the probable error of the mean of each sample. In the first sample the mean is 12.40, the standard deviation of the sample is 3.21, and the probable error of the mean is .48. The mean is then written as $12.40 \pm .48$, which indicates a probable error range from 11.92 to 12.88.

Interpretation of the probable error. We now come to the interpretation of the probable error range. Just what does it mean? Once in a while the above statement of the average and its probable error is interpreted by saying that the true mean is necessarily between 11.92 and 12.88. This is wrong. The true mean of the large pile of cards is 12. The cards were so arranged, and the true mean was determined to be exactly twelve. In this particular sample the true mean does lie between the probable error limits 11.92 and 12.88. The answer to this question is "yes" and it is so labeled in column five. In the next sample the mean was found to be 13.15 with a probable error of .64. The probable error range for that sample is from 12.51 to 13.79, and since the true mean is at 12, it is seen that the probable error range

of the mean of the sample as well as the mean itself of the sample are entirely beyond the true mean.

We have the question "Is the true mean inside the probable error limits?" This question is answered "yes" 26 times, and "no" 24 times. This leads to the inference that the true mean of the universe from which the samples are drawn may be either inside or outside of the probable error range of the sample. In fact, the chances are even that the true mean is inside the probable error limits of any single sample. But it is important to remember, and this is what the student occasionally forgets, that the chances are even that the true mean is *outside* the probable error limits of the sample. While the chances are even that the true mean is inside the probable error limits, the chances are also even that the true mean lies outside of these limits. When we say this, we are dealing only with the fluctuations in the average that are caused by mere chance, and we know nothing whatever, as far as inference from the probable error is concerned, regarding any illegitimate or unintentional loading of the sample which may misrepresent the unmeasured total.

Suppose that we are measuring the stature of ten-year-old children. We obtain records of stature for twenty ten-year-olds. We calculate the average stature for this group and we determine its probable error. If the probable error is small, we regard the determination of average stature for ten-year-old children to be fairly accurate. We must remember, nevertheless, that the chances are even that the true

height of ten-year-old children is *outside* of the limits set by the probable error of the average stature.

The practical situation is parallel to this experiment with the fundamental exception that we do not know anything at all about the *universe*, all ten-year-old children in the city or district where we are working. This universe is represented in the experiment by the large pile of cards. Under actual working conditions in which statistics are applied we know nothing about the total that we are trying to measure. We only know about the records in one single sample which may contain twenty cases, or one hundred cases, or several thousand, as the importance of the job and available records may dictate. We are always working with a sample, just as fruit, or cement, or grain is evaluated by samples. If the sample is taken out of the carload, or the bag, entirely at random, it will be more or less representative of the entire carload, or the whole bag. It will be subject to chance variation from one sample to the next, and such variation can be predicted as to its extent by the probable error formulæ if the evaluations are quantitative. But the sample may have been drawn with an intentional or an unintentional bias such as would be the case if conditions of exposure to heat, light, and moisture in the carload were to affect the sample in a way which would not represent the conditions in the entire car. That would be an unintentional bias. The top of a fruit basket constitutes a sample which is intended to serve as representative of the entire basket, and so the consumer learns what is

meant by intentionally loaded samples. These different kinds of bias have their logical counterparts in the samples of statistical work in education and in the social sciences. The probable error tells us something regarding the reliability of our estimates of a total from the relatively small sample by which we try to prove something regarding the total. But the probable error tells us only about the expected variations that are due to mere chance, and these chance variations are very small compared with the misrepresentations that are brought into statistical work by intentional and unintentional bias in the sample. About the more serious factors that disturb statistical validity, the probable error tells us nothing.

If, instead of knowing all about the large pile of cards, we decide to take a sample of twenty cards by which to estimate the characteristics of the total, we shall have one average for our sample, and we shall be obliged to make the best possible estimate of the reliability of that average in describing the unmeasured total. We shall have no means of knowing which one of the fifty or more samples it is that we have drawn. We may just happen to be working with one of the samples that has a probable error range entirely outside of the true mean. The fact that the true mean is at twelve is not known because in practical statistical work the total is never known. All we can do is to determine the average of the particular sample with which we are working, calculate the probable error of that average, and then state that there is an even chance that the true average

is *outside* of these probable error limits. Further, we give assurance that the still more serious factors that might cause a sample to misrepresent the total have been avoided not by figuring the probable error but by selecting the members of the sample as far as possible at random and unaffected by any time and place characteristics which are not a part of the definition of the total that the findings are supposed to describe.

Definition of probable error. If we have one sample of twenty cards with its average and its probable error, can we say anything regarding the corresponding facts for the next sample of twenty cards that we might draw if opportunity were to present itself? For example, can we be fairly sure that the average of the next sample that we might draw will fall between the probable error limits of the first sample that we already have drawn? No. If we consult the records in the experiment as shown in the table, we shall find that in approximately half of the cases the average of the next sample will be inside the known probable error limits, and in approximately half of the cases the average of the next sample falls outside of these limits. Therefore, we can say with fair assurance that if the next sample be drawn under conditions that are absolutely parallel to the conditions obtaining for the first and completed sample, a situation that rarely obtains in practice, there is an even chance that the average of the next sample will be outside of the probable error limits that we already have. This principle can be verified in the table.

The probable error (denoted by P. E. or E.) of a mean is a pair of divergencies, lying one above and the other below the mean, of which one can say with confidence that there is an even chance that the true mean lies between these limits. This definition of the probable error assumes that noted deviations are all due to chance.

If we desire to specify a range, with more certainty that the true value lies within the specified range, we may do so by the aid of the following table. The chances that the true value lies within the range set by $\pm E$, $\pm 2E$, etc. are as follows:

E — the chances are even
$2E$ — the chances are 4.5 to 1
$3E$ — the chances are 21 to 1
$4E$ — the chances are 142 to 1
$5E$ — the chances are 1310 to 1
$6E$ — the chances are 19,200 to 1
$7E$ — the chances are 420,000 to 1
$8E$ — the chances are 17,000,000 to 1
$9E$ — the chances are about 1,000,000,000 to 1 [1]

If we apply these probabilities to the average and its probable error for the first sample of twenty cards in *Table 16*, we find that the mean is 12.40, that the probable error is .48 and that twice the probable error range would be from 11.44 to 13.36. We can therefore say that the chances are 4.5 to 1 that the true mean is between 11.44 and 13.36. It should be clear that the more we extend the range the greater will be the probability that the true mean is within it.

[1] C. B. Davenport, "Statistical Methods," p. 14.

The difference between the probable error and the quartile or median deviation. In *Figure 35* we have two frequency distributions, one distribution for the original separate numbers and one distribution for the averages. The small dots represent this latter frequency distribution. Each dot represents an average of twenty numbers. The spread or variability of this distribution is much smaller than the variability of the original numbers. The probable error of the mean of a single sample is comparable with the semi-interquartile range of the distribution of averages. One-half of the averages falls within the probable error limits, and one-half of them falls outside, just as is the case with the quartile deviations of any distribution.

The terms *probable error* and *quartile deviation* are statistically similar, and their numerical values are identical, but the terms are used in different ways. If we are talking about the spread of a lot of numbers that we have actually recorded, we measure the known spread by the quartile deviation. If we are estimating or predicting, from a single sample, what the spread would be, we call the same idea the probable error. If we have a sample of twenty numbers, and if we have calculated the average and its probable error, we mean by the probable error an estimate or prediction of expected variation in the mean if additional samples were taken. Now suppose that, instead of estimating what the variation would be, we actually proceed to take additional samples and calculate the mean for each one. Suppose that we

have done this for fifty or one hundred samples. We can then plot a frequency distribution of the averages. If we do that, the probable error would be referred to as the quartile deviation of the distribution of averages. The two concepts are identical but they are used to designate two different situations. *If we actually have the distribution, we measure the variability and call it the quartile deviation. If we are merely estimating what the variability might be, we call it the probable error.* Both concepts denote the same measure. They are both laid off from the best obtainable mean, and they both include within their range one-half of the measures in the distribution.

Some applications of the probable error formula. Under certain conditions the following type of question might arise: What is the probability that the true mean is not less than 12? If we have the first sample with its mean of 12.40 and if we do not know the true mean, we could make an estimate of the probability of the true mean being as low as 12. Supposedly the value 12 is .40 below the mean we obtained in the only sample of which we have any knowledge. The question is then to determine the probability of the true mean for the unmeasured total being .40 below the mean that we actually did obtain in the one sample. The probable error of the obtained mean (12.40) is .48 and hence the distance .40 can be expressed as a deviation of $\frac{.40}{.48} = .83E$. If we should plot a frequency curve about 12.40 as a mean, with a quartile deviation of .48, we should

have the best possible guess as to what the distribution of averages would look like if we were to proceed with a lot of additional samples like the one that we have actually taken. The true mean about which this distribution of averages should be plotted is of course not known, but we use the mean of the one sample as the mean of the expected distribution of averages because it is the only average of which we know anything. The chances are even, as we have seen, that this obtained average for a single sample is above the true mean, and the chances are also even that this obtained average is below the true mean of the unmeasured total.

If we imagine this expected distribution of averages of additional samples that we might take, and if we consider the fact that the one sample already available might be any one of the points in that distribution, our question of ascertaining the probability that the true mean is not lower than 12 can be restated in the following form. What is the proportion of the area in the distribution of expected averages which lies above 12? The point 12 is designated as $-.83\ E$. The standard deviation and the probable error are always in the following relation: $P.E. = .67\ \sigma$. Hence the point 12 is designated on the sigma scale as $-.56\ \sigma$. The proportion of the entire surface between the mean ordinate and this point $-.56\ \sigma$ is 21% and hence the probability of the true mean being not less than 12 is .71. The probability that the true mean is less than 12 is therefore .29. It so happens, as we discovered by actu-

ally taking the additional samples that we are here guessing about, that the true mean is at 12.

Another type of problem that involves the use of the probable error concept would be the following: What is the probability that the true mean does not differ from the mean of the sample by more than, let us say, .3? We start as before. The only average that we know anything about is the average of the one sample of twenty cards, which is 12.40. Our question can be restated in the following form: What is the probability that the true mean of the unmeasured total lies between $12.40 \pm .30$ or between 12.10 and 12.70? We imagine again a frequency distribution of the averages of the additional samples that we might take, similar to the one that we have already drawn. We assume that the mean for this distribution of averages is at 12.40 because that is the only average about which we know anything. The probable error of this average is .48 and hence the range that we are now interested in can be expressed as the range from $-.62\,E$ to $+.62\,E$ because the deviation .3 is .62 of the probable error of the mean, .48. The range thus expressed can be turned into a sigma range from $-.42\,\sigma$ to $+.42\,\sigma$. The proportion of the whole surface between these two limits is, according to the appropriate tables, 33%. Hence the probability is .33, that the true mean is between the limits 12.1 and 12.7. As a matter of fact the true mean is 12, but that is of course not known when we are making these estimates on the basis of a single sample.

Still another application of the probable error concept would be to the following question: What is the range, above and below the mean of the sample, that we must allow in order to be able to say that the chances are 10 to 1 that the true mean is between the limits so given? We shall make the application again to the first sample of twenty cards. The mean of that sample is 12.40 and we imagine a distribution of averages about this mean. The probable error of the mean is .48 and that becomes the quartile deviation of the expected distribution of averages. We now want to find how far we must travel from the mean in both directions so that we shall include $\frac{10}{11}$ of the surface and leave $\frac{1}{11}$ of the surface above and below these limits. The fraction $\frac{10}{11}$ is .91; half of this will be above the mean, and the other half below the mean. We find, from the appropriate table of fractional areas, that the required 45% of the distribution lies between the mean ordinate and the point 1.65 σ or 2.46 E. Since the probable error is .48, we locate the limits of the range as 2.46 × .48 or 1.18 above and below the obtained mean of 12.40. The range required is therefore from 11.22 to 13.58 and we can say that the chances are 10 to 1 that the true mean is between these limits. Do not forget that, as far as one can know from the single sample with which we are working, the chances are still $\frac{1}{10}$ that the true mean is outside of this range, and we may just happen to have that one sample which we estimate would occur once for ten samples of the other kind.

The standard deviation as a measure of reliability. We have used, as measures of variability, not only the quartile deviation but also the standard deviation. Both of these measures of actual variability may be used for estimating expected variability. We find, therefore, in statistics not only the probable error but also the standard deviation used to indicate expected variation. Since the standard deviation, when laid off from the mean in both directions, includes two-thirds of the entire surface, and since the quartile deviation, or probable error, when so laid off from the mean, includes only one-half of the measures, it follows that the standard deviation, when used as a measure of reliability for an average or other constant, will indicate a larger range from the obtained mean. The interpretation of $12.40 \pm .72$, where the .72 indicates the standard deviation of the mean, would be that the chances are two to one that the true mean is between the limits so given. All of the calculations and interpretations of the standard deviation as a measure of reliability are similar to those of the probable error, with the exception that the standard deviation from the mean marks off two-thirds of the entire surface, whereas the probable error when so laid off includes one-half of the cases.

The probable error of other statistical constants. The probable error, or expected quartile deviation, is used as a measure of reliability for other statistical constants such as the median, the standard deviation, and the Pearson correlation coefficient. The interpretation is the same. Thus, if we calculate the

probable error of a Pearson correlation coefficient, we might express it as $+ .62 \pm .05$, which would mean that the chances are even that the true correlation is between .57 and .67.

It will have been seen that the probable error is calculated on the assumption that the distribution in the sample is normal. When the distribution is not normal, and especially when it is noticeably asymmetrical, the probable error loses its customary significance. As long as the distribution of the sample is bell shaped, and not considerably skewed, the probable error may be applied to its constants with fair assurance of its applicability.

Probable error formulæ.

Arithmetic Mean: $P.\ E. = .67449 \frac{\sigma}{\sqrt{n}}$,

in which σ represents the standard deviation of the sample, and n is the number of cases in the sample.

Standard Deviation: $P.\ E. = .67449 \frac{\sigma}{\sqrt{2n}}$,

in which σ represents the standard deviation of the sample, and n is the number of cases in the sample.

Pearson Coefficient of Correlation (see page 205):

$$P.\ E. = .67449 \frac{1 - r^2}{\sqrt{n}},$$

in which r is the correlation coefficient, and n is the number of cases.

The corresponding standard deviations may be determined from the relation $P.\ E. = .67449\ \sigma$.

Summary. We have seen how the quartile deviation is a measure of the variability or scatter of the numbers in a distribution. If, instead of dealing with the original numbers and their scatter, we are dealing with the scatter of averages of successive samples, we can apply the same measures of variability. Ordinarily we have only one sample, which may contain twenty numbers or two thousand numbers, as the case may be. For that one sample we obtain an average. In estimating the reliability or confidence that we should attach to this average we imagine that additional samples are taken and the average calculated for each one. Of course, if we have a large number of cases in each sample, there will be relatively small fluctuation in the successive averages. If the numbers in each sample cluster very closely about the average of the whole sample, we can be fairly certain that the successive averages would fluctuate less than if the numbers in each sample were found to scatter widely over a long range. These two factors, the scatter of the numbers in any one sample and the number of cases in the sample, go to determine the probable error or reliability of the average of the sample. If we actually draw these additional samples, as we have done in the experiment here described, and if we plot the distribution of averages, we may actually measure the extent to which they scatter by determining the quartile deviation of the distribution of averages. If, instead of actually drawing successive samples and actually plotting the scatter of the successive aver-

ages, we estimate what the fluctuation in successive averages would be, on the basis of a single sample, then we talk about the expected quartile deviation as the probable error. A quartile deviation which is not actually measured, but only estimated or predicted, is called a probable error.

Problem 1. The arithmetic mean of a sample is 64.78; its standard deviation is 6.34 class intervals; each class interval contains .5 scale units. There are 100 cases in the sample.

1. Determine the probable error of the mean. Of the standard deviation.

2. Determine the standard deviation of the standard deviation.

3. What is the probability that the true mean is higher than 68? That it is less than 64.78? That it lies between 64 and 65?

4. What range must be specified on both sides of the obtained mean so that one may say that the probability is as five to one that the true mean is within the range?

5. What is the probability that the true mean does not differ from the obtained mean by more than .3?

6. If 150 additional samples were taken similar to the above sample, and the average calculated for each one, what semi-interquartile range would you expect for the distribution of the additional 150 averages? What would be the expected semi-interquartile range of the averages of 300 additional samples with the same number of cases in each sample?

7. Assume that the true mean is at 66. Plot the expected distribution of the 150 additional averages. Indicate on the chart the location of the average for the above sample.

Problem 2. Write an explanation of what is meant by the standard deviation of the standard deviation. Explain how you would arrange an experiment to verify your definition empirically.

Problem 3. The following problem should afford some good points for discussion.

Suppose that you have determined the average or mean stature for a random sample of 50 students from a class of 500 and that you have calculated the probable error of this mean. Suppose, further, that you have calculated the mean stature for the 50 tallest men in the class and its probable error. Which probable error is the smaller? Can you explain the reason?

Problem 4. Assume that two cities, *A* and *B*, have populations of 1,000,000 and 1,000, respectively, and that you have drawn a sample of 500 cases from each population to determine average stature, age, or any other measurement. Which of the two means is the more reliable index of the population that it represents?

Is the mean of a random sample of 500 affected by the size of the population from which it is drawn?

Is the probable error of the mean of the sample affected by the size of the population from which it is drawn?

Is the mean of the sample more reliable if it contains one-half of the population which it represents than if it contains only $\frac{1}{2000}$ of that population? How does this inconsistency come about?

Is the reliability or trustworthiness of the mean of a sample of 500 increased if the sample contains the whole population? Is this factor shown by the probable error formula?

Discuss the limitation in the meaning of the probable error for those situations in which the population or universe is reduced to a number comparable with the sample itself.

Chapter Twenty-two

The Correlation Table

Relation between variables. Every scientific problem is a search for the relationship between variables. Every scientific problem can be stated most clearly if it is thought of as a search for the nature of the relation between two definitely stated variables. Very often a scientific problem is felt and stated in other terms, but it cannot be so clearly stated in any way as when it is thought of as a function by which one variable is shown to be dependent upon or related to some other variable. Statistical discussion sometimes becomes unnecessarily muddled by the fact that the investigator has not stated clearly for himself just what the several variables are that he is studying, and just what the units are by which he is measuring and describing them. This cause of fog is most serious in scientific discussion that is limited to words. When a scientific inquiry is not quantitative, when it is carried out entirely in terms of language, it becomes especially desirable, for the sake of clarity, to state what the variables are, and to state in what terms these variables are described. The core of the problem is always found to be in the nature of the interdependence of the two or more variables. When that interdependence or relationship has been clearly stated and verified, the scientific problem has been solved.

The only exception, a partial one, may be the initial stages of scientific work in a virgin field, in which scientific labor is largely descriptive. At that stage in the exploration of a novel field of facts the curiosity of the investigator is largely directed toward the description of unknown things. Sooner or later, however, he or his followers reach the more truly scientific aspect of inquiry that consists in establishing dependable relations between the facts found or between the new facts and the facts that are already known. An archæologist who has just discovered an Egyptian tomb containing mummies and other relics has first the task of describing what he has found. Such labor is part of his scientific work, but it is only preliminary to his real contributions to science in relating the discovered facts to one another and relating them in turn to facts of history already established.

A physicist who is investigating the velocity of sound through different substances is describing them. He is listing a quantitative measure of this attribute for each substance. The more truly scientific aspect of his problem appears when he becomes curious about the relation between the velocity of sound through the substance and its other known characteristics, such as density, chemical ingredients, the chemical history of the substance, boiling point, electrical conductivity, or shape. He may then discover that some of these characteristics have no relation whatever to the velocity with which the substance conducts sound, while other characteristics do

have such a relation. It is in the discovery of such interrelations that we have the basis for prediction and control which constitute the practical applications of science.

Correlation statistics are concerned with the relation of measured variables. We have so far been concerned primarily with the characteristics of a single variable such as the central tendency and the variability. We shall now consider the corresponding statistical methods for studying the relation between variables.

The logic of the correlation table will be clearer, perhaps, if we first note a few examples of scientific problems. It is a scientific problem to ascertain the strength of different materials of construction. A beam will sag if a load is applied to its middle. The particular problem may be to determine how much it will sag under different applied loads. Here the two variables are the load, measured in pounds, and the amount of sag, measured in any suitable units of length. In this case there is found to be a direct relation between the two variables, because as one of the variables increases the other one increases also. It would also be discovered that the shape or cross section of the beam determines the amount of sag. In such a problem the two variables might be the depth of the beam and the amount of its sag. It is found that if the depth of the beam is increased, the sag decreases, provided one is experimenting with the same load for the different trials. The relation between these two variables is therefore inverse. The

load would have to be kept constant in the experiment in order to obtain the true relation between the shape of the beam and its sag. As soon as a problem is stated in scientific form two or more variables are involved and our goal is the establishment of the nature of their relation.

If we are investigating the barometric pressure and its usefulness for predicting the weather, we have a scientific problem in which one variable is the barometric pressure and the other variable is the related amount of precipitation, or sunshine, or wind. As soon as we state a scientific problem definitely two or more variables appear. Our thinking about the problem and the technique for solving it becomes clearer when we have clearly stated just what our variables are.

A *variable* is anything that can have a series of values, such as height in inches, age in years, sag of the beam in millimeters, air pressure in mercury centimeters, income in dollars. We have seen that variables can be dealt with in statistics conveniently by sorting the individual measures into class intervals. For this purpose we ordinarily select ten or twenty or more intervals or successive classifications. In handling certain variables it is found that, for practical purposes, there are only two class intervals. If we have the stature of one thousand men, we may classify them into two class intervals, those who are above the average and those who are below it. These two class intervals would be designated tall and short. When we see any attribute classified into

two groups or classes, we may still recognize that the attribute is in the nature of a variable and that the classification into class intervals is roughly made into only two parts instead of the more customary fifteen or twenty class intervals. Consider the scientific problem of ascertaining whether mosquitoes have anything to do with malaria. Here the two variables are the frequency of exposure to mosquitoes and the incidence of malaria. These are the two variables of the problem, both of which might conceivably be stated in numerical form. The frequency of exposure to mosquitoes could be handled by comparing two groups of men, one group being exposed frequently to the bite of mosquitoes and the other group being entirely shielded from them. Such an experiment would show a difference in the incidence of malaria and so the relationship between the variables would be established, although for experimental purposes one of the variables would be represented by only two of its extreme "class intervals." A chemical problem of the same type, statistically, would be that of identifying a metal by qualitative analysis. One of the variables is in that case the amount of the metal present in the sample. It is represented in an experiment by its two extremes, the absence and the definite presence of the metal in the sample.

Independent and dependent variables. In almost every practical situation in which the result of scientific work is to be used there is a distinction between what is known as the independent variable and the dependent variable. If we are studying the relation

between barometric pressure and the weather, as measured by sunshine, humidity, wind velocity, precipitation, there is usually a variable that is known and that is used for predicting the variables that are unknown. The variable that is ordinarily known first is called the *independent variable*. The variable that is being predicted or inferred is called the *dependent variable* because it is somehow dependent on the facts that are already known. The barometric pressure would, in weather prediction, be called the independent variable and the measurements of the resulting weather conditions would be called the dependent variables.

If we are studying the relation between the shape of a beam and its sag, the shape of the beam would be called the independent variable. The sag, being dependent on the shape of the beam, is known as the dependent variable. The nature and extent of the sag depend upon the shape of the beam.

If a college entrance examination is to be evaluated against college scholarship as a criterion, the entrance examination mark would be the independent variable and the college grades would be the dependent variable.

Direct and inverse relationship. If the relation between two variables is such that as one of them increases the other one increases also, the relation is said to be direct or positive. If, on the other hand, as one of the variables increases the other one decreases, the relation is said to be inverse, or negative. If an educational test in arithmetic is given to

a group of students, and if the mark in the test consists in the amount of time consumed to finish a predetermined number of problems, it is clear that the longer the time, the greater will be the score which is measured in minutes, and the lower will be the arithmetical ability of the student. The relation between the mark in the test and the trait which it is supposed to measure is therefore inverse or negative. If, on the other hand, the test score consists in the number of problems solved in ten minutes, the marks would increase as the abilities of the subjects increase, and hence this relation would be direct or positive.

Degree of relationship. It is possible for two variables to be related positively and yet imperfectly. There may be a general tendency for one variable to increase as the other one increases and yet the relation may not be sufficiently uniform or close to make it possible to assert it for each individual case. An example is the relation between height and weight. In general tall people weigh more than short people. There is a direct and positive relation between height and weight, but it is not perfect. When several relations are being compared, it happens frequently that of two relations that are both positive and imperfect, one may be closer than the other. The degree of relationship is the central topic of correlation statistics. We shall describe several examples showing various degrees of relationship from perfect positive to perfect negative.

There is an almost perfect relation between the weight of a piece of steel and its volume. The weight

is practically proportional to the volume. This relation is positive and perfect. The relation between height and weight is also positive, but it is not nearly so close and dependable as the relation between the weight and the volume of a piece of steel. The height-weight relation is positive, but imperfect. A relation that is practically zero is the relation between a man's income and the size of his shoes. If there is any relation between these two variables, it may be very slightly positive. The relation between the cubic weight of a substance and its money value is probably zero. The relation between the time consumed in doing a piece of work and proficiency in doing that work is negative but imperfect. An example of a negative relation that is almost perfect is the relation between the volume and the pressure of gas when the temperature is kept constant. As one of these variables increases the other one decreases.

Relations in the social sciences. The study of relationships in the exact sciences is ordinarily made in quantitative form, and the methods for doing it are so common that they are almost taken for granted by physicists, astronomers, engineers, and others who are dealing with quantities that can be measured accurately. When a physicist has a mass of data to be examined, he naturally transfers the records to a chart in order to discover new functional relations between variables. In the biological and social sciences these quantitative methods for discovering relations are not so generally used because the vari-

ables are not so frequently susceptible of quantitative measurement.

With the advancement of any science there appears an increasing amount of quantitative work. This is becoming evident in several of the biological sciences, including psychology. With development, the social sciences tend to become biological sciences, and these in turn are destined to become exact sciences.

All the sciences deal with variables. There is, however, among the sciences considerable difference in the precision with which the variables may be isolated for measurement. The physicist is able to isolate most of the irrelevant factors or variables. He obtains data showing the relation, or absence of relation, and the exact nature of the relation. In the social and biological sciences it is not possible to isolate the irrelevant factors or variables. For example, an economist may be studying, as a scientific problem, the relation between density of population and urban rent. He can obtain accurate records of rent, and he can obtain records of the number of people who are working or living in each city block. But he is not able to do what the physicist can do in the laboratory, namely, to shut out the effect of factors that he is not studying. The economist cannot regulate for experimental purposes the many other factors that partially determine rent, such as transportation facilities, elevation of ground, proximity to shops and parks, prestige of the district, and age of the buildings. He must deal with the records as he finds them, influenced and partly determined by each

of many factors, the exact part played by some factors being unknown. It is possible to determine tendencies in the relations by accumulating a sufficient number of records, but the charts will show considerably more scatter of the data than is found in the charting of observations in the exact sciences.

The correlation methods in statistical work have been developed to facilitate the study of relations between variables where the records show considerable scatter. The methods that are used in the exact sciences are slightly different in appearance but they can be shown to be mathematically identical. Thus the physicist describes his data by means of an equation, the constants of which have been determined by the method of least squares. The economist and the psychologist describe their data by means of a regression equation and a correlation table. It can be shown that the equations are absolutely identical, although the routine methods of presentation are slightly different on account of differences in the customary amount of scatter that is found in the observations of the respective sciences.

The scatter diagram. The scatter diagram is a chart for showing graphically the relation between two variables. The scatter diagram shows graphically not only the presence or absence of relationship, but it also enables one to judge by inspection the degree of relation between the two variables plotted. In *Figure 36* there are five scatter diagrams showing the appearance of these diagrams for different degrees of relationship. In *Part 1* we have a scatter diagram

showing a perfect positive relation between two variables such as the volume and weight of pieces of steel. Every member of the group is represented by a point in the scatter diagram. The members of the group may be persons whose stature, weight, intelligence, or salary are being plotted, or the members may be any material things about which two variables are being studied, such as the pressure and volume of a quantity of gas, in which case the pairs of observations of the same quantity of gas would constitute the members of the statistical group. If the

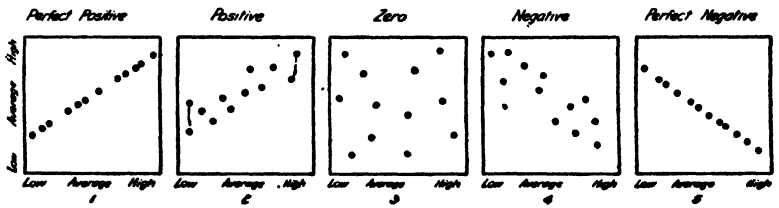


Figure 36. Positive and negative scatter diagrams

relation between rents and density of population is being studied, the members of the statistical group would be the sections or city blocks or the buildings, for each one of which two variables or facts are noted. Every member of the group is represented by a point on the scatter diagram.

Every point on the scatter diagram tells *two* facts. The distance of the point from the horizontal base line of the chart tells one fact, the measurement of one of the variables. The distance of the point from the left vertical edge of the chart shows the second fact, the measurement of the second variable. Some-

times it is not feasible to start the measurements on the chart with zero. In that case the two facts are to be read off from the two scales of the chart.

In *Part 1* of *Figure 36* we have the x-variable indicated on the base line running from left to right, from low values to high values. The y-variable is indicated on the vertical axis, running up on the chart from low values to high values. It can be seen at a glance on this chart that as the values of x increase the values of y also increase by a strictly proportional amount. The relation is therefore perfect and positive.

In *Part 2* of the same figure we have a scatter diagram with twelve paired observations for a relation that is positive but imperfect. It is clear by inspection of *Part 2* that as the values of x increase the values of y tend to increase in general, but the increase in y is not for every observation strictly proportional to the increase in x. Thus there are two observations which indicate different values for y at the lowest value of x, and there are similarly two different values for y at the highest value of x. This shows that the relation is not absolutely dependable and exact, but the average y-value for the six highest x-readings is higher than the average y-value for the six lowest x-readings.

In *Part 3* we have a scatter diagram which shows entire absence of relation between two variables. Here we see that the average y-value for the six highest x-values is about the same as the average y-value for the six lowest x-values. To change the value of one of these variables does not have any effect

on the other variable even if we consider the general tendency. The relation is said to be zero or absent.

In *Part 4* of the same figure we have a scatter diagram for a relation that is imperfect and negative. It is negative because as we increase the values of x the values of y tend to decrease. Note that the average y-value of the six highest x-values is lower than the average y-value of the six lowest x-values. As one of these variables increases the other one decreases. An example would be the number of arithmetic problems solved per hour and the average time per problem. As one of these variables, the number of problems, increases, the other variable, time per problem, decreases. The relation is negative. In the scatter diagram of *Part 4* the relation is imperfect.

In *Part 5* of the same figure we have a scatter diagram of a relation that is perfect and negative. As one of the variables increases the other one decreases and the relation is such that the changes in the two variables are strictly proportional to each other. A relation may be perfect even though the changes in the two variables are different. One of the variables may have numerical changes covering hundreds of its scale while the other variable has changes covering only tenths of its scale. There is no necessary relation in the magnitude of the two scales, but in a perfect relation the changes are strictly proportional throughout the entire range for which paired observations have been plotted.

We have seen in dealing with a single variable that

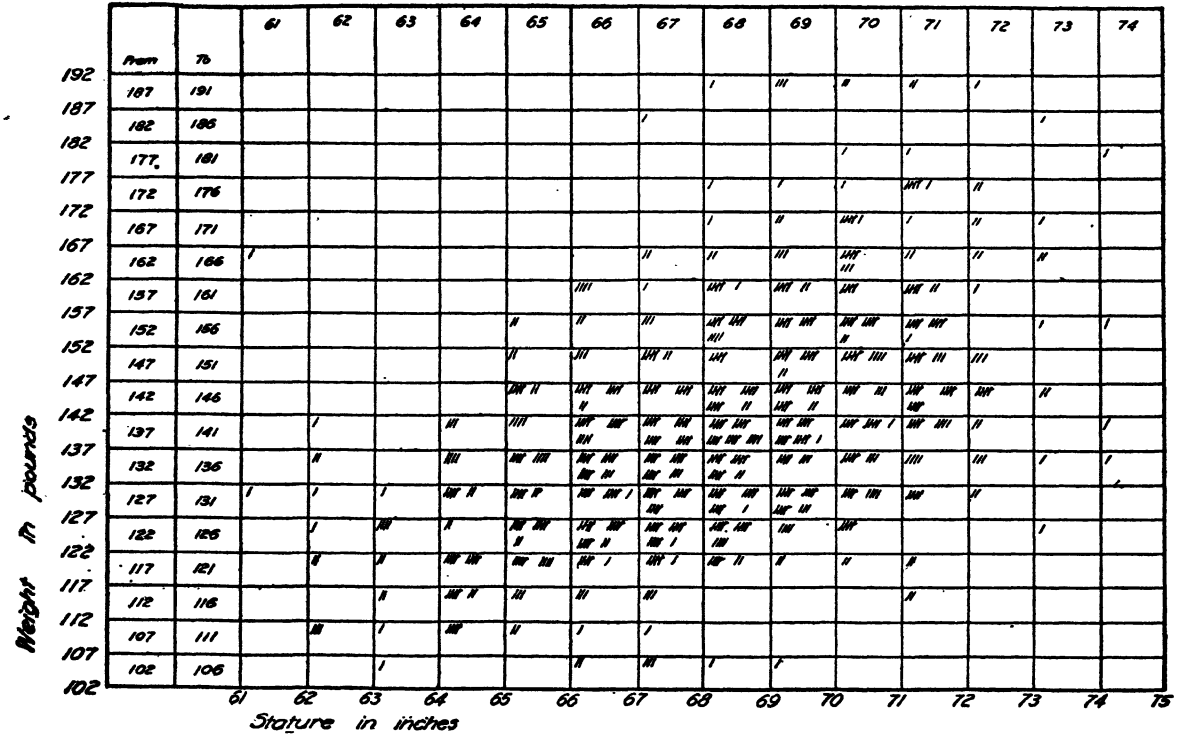


Figure 37. Scatter diagram for height and weight

it is convenient to divide the scale of the variable into class intervals for purposes of tabulation. We shall do likewise for plotting scatter diagrams. In *Figure 37* we have a scatter diagram for the relation between height and weight of 750 students at Ohio State University. Every person in the statistical group is represented by a mark in one of the squares. Since there are 750 men in the group there are as many small marks on the diagram. Every mark represents two facts about a student. His weight in pounds is indicated by the scale at the left edge of the diagram. His stature in inches is indicated by the scale along the bottom of the diagram.

In *Figure 38* we have the corresponding data for the relation between the height and the weight of 750 students plotted in the form of a correlation table. The only difference between a scatter diagram and a correlation table is that in the scatter diagram we have a point for every case, whereas in the correlation table we have the total frequency in each box listed in numerical form. Otherwise the scatter diagram and the correlation table are identical. It is of course necessary to prepare something corresponding to the scatter diagram in order to be able to count up the frequencies in the boxes of the diagram.

If we look at the correlation table in *Figure 38*, we find that the average weight tends to increase from any class interval in height to the next higher class interval in height. This is apparent by the diagonal arrangement of the frequencies. The upper left corner and the lower right corner of the diagram are

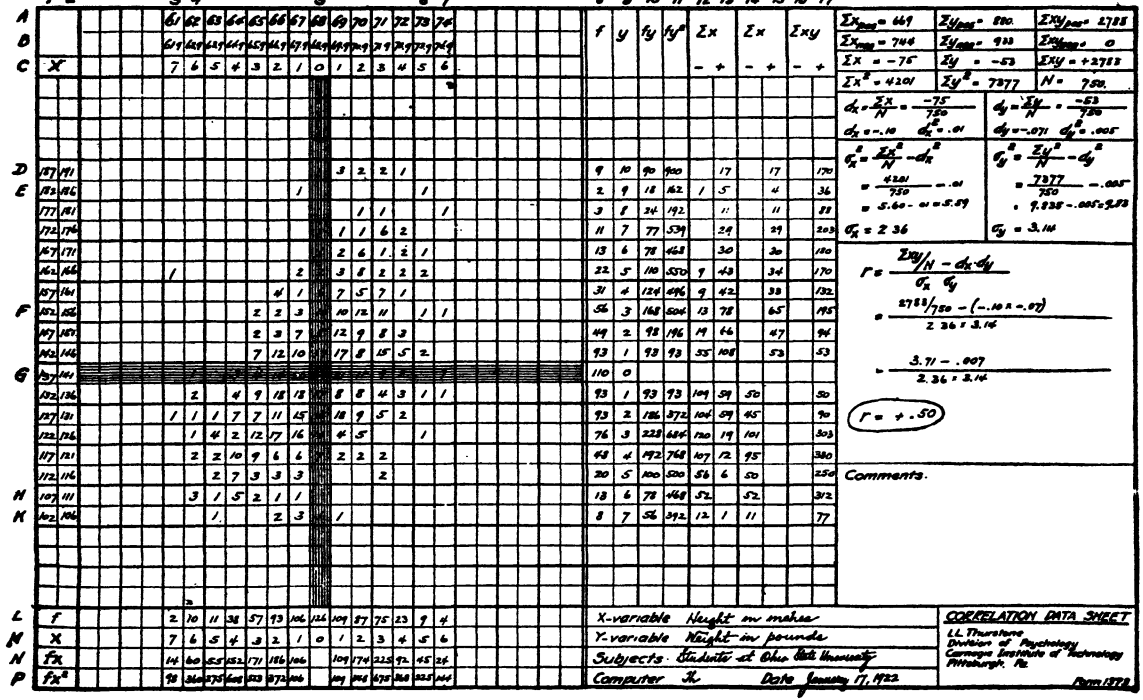


Figure 38. Correlation data sheet[1]

[1]The blank forms of this correlation data sheet are distributed by C. H. Stoelting & Co., 3047 Carroll Ave., Chicago, Ill.

relatively empty, whereas the diagonal area from the lower left corner to the upper right corner of the diagram is relatively well filled. The extent to which the marks or frequencies in the diagram cluster around a diagonal line indicates roughly the degree of relationship between the two variables.

If there were no relation between the two variables, the diagram would have an entirely different appearance. If, for example, we should tabulate the relation between each man's income and the size of his shoes, we should probably find no noticeable relation. The frequencies for income would come as high in the diagram opposite the small sizes of shoes as opposite the large sizes of shoes. The points or frequencies in the diagram would cluster around the center and the four corners of the diagram would be about equally filled. There would be no noticeable diagonal arrangement and consequently no relation would be observed.

For many practical purposes it is sufficient to plot the correlation table or the scatter diagram and to analyze directly from it the desired information without calculating the correlation coefficients and the regression coefficients. From the correlation table one may extract all the information necessary for plotting bar diagrams, curves, and other charts, and for tables of averages. It should be borne in mind, however, that the correlation table gives all the facts about the relation between two variables, and that all other forms of representing statistical results give only a part of the evidence.

Problem 1. Prepare a chart to show the average weight for students whose stature is 61, 62, 63 inches, etc. (See *Figure 37.*)

Problem 2. Prepare a chart to show the average stature for students whose weight is 105, 110, 115 pounds, etc. What is the average stature for students of your weight? What is the average weight for students of your stature?

Problem 3. Prepare two tables showing the average statures of students for given weights, and the average weights for students of given statures.

Chapter Twenty-three

The Pearson Correlation Coefficient

The *correlation coefficient* is a pure number, a constant which indicates the degree of relation between two variables. It varies from + 1 to − 1. When the relation is perfect and positive, the correlation coefficient is + 1. When the relation is perfect but inverse, the correlation coefficient is − 1. When there is no relation whatever between the two variables, the coefficient is zero. Other values of the coefficient indicate intermediate degrees of relation. Thus a coefficient of + .8 indicates that the points on the scatter diagram cluster rather closely about a diagonal line across the diagram, whereas a coefficient of + .3 indicates that the points scatter more from the diagonal tendency although the relation is still noticeable. The degree of relation between height and weight is approximately + .5.[1] It is apparent, then, that the correlation coefficient is only a numerical way of describing the scatter diagram, although the diagram gives more information than can be found from the single numerical value of the coefficient. When a great number of relations are being studied, the correlation coefficients serve as abbreviations or indices of the degree of relation from which

[1] Students occasionally make the mistake of calling a correlation coefficient a per cent. A correlation of + .5 is called "five tenths" or "fifty." It is not a per cent. It is a pure number.

the experienced statistician can visualize the diagram, more or less roughly. If one has the option of seeing the scatter diagram and the correlation coefficient, one would of course choose the diagram because the coefficient can be found from the diagram but the diagram cannot be at all accurately constructed from the coefficient. The diagram gives more information than the coefficient, but when many relations are to be compared, the coefficient serves as an objective and impartial measure of the degree of relation.

The Pearson coefficient of correlation is usually found by the following formula:

$$r = \frac{\Sigma xy}{n \cdot \sigma_x \cdot \sigma_y},$$

in which

x = the deviation of the x-values from the mean of x
y = the deviation of the y-values from the mean of y
n = the number of cases in the diagram
σ_x = the standard deviation of X
σ_y = the standard deviation of Y
r = the Pearson coefficient of correlation

It is apparent that, in order to calculate the coefficient from this formula, it is necessary to determine the product of the xy deviations for each point, the number of cases, and the two standard deviations.

It is good practice for the beginner to calculate the correlation coefficient for a diagram of about fifty cases with the actual deviations from the mean with-

out using an arbitrary origin, and to do likewise for the two standard deviations. In actual practice, however, it is convenient to use a data sheet specially prepared for the calculation of the coefficient. The purpose of a data sheet is to facilitate the computations and to provide a definite space on the sheet for each sum or product.

The correlation coefficient can be calculated without using the deviations from the means.[1] The formula often leads to unusually large numbers which are awkward to handle. They may be corrected by using equivalent scales instead of the original X and Y numbers. The formula is as follows:

$$r = \frac{\Sigma(XY) - n \cdot m_x \cdot m_y}{\sqrt{\Sigma(X^2) - n \cdot m_x^2} \cdot \sqrt{\Sigma(Y^2) - n \cdot m_y^2}},$$

in which m_x = mean of the x-values

m_y = mean of the y-values

It is well to take into consideration the several factors that influence the size of the correlation coefficient. If the observations are themselves inaccurate, it is of course evident that the correlation coefficient may be lower than it would be if the observations were accurate.

It sometimes happens that two variables are related through a common third variable which, if not controlled or kept constant, plays havoc with the

[1] For a more detailed description, see L. L. Thurstone, "A method of calculating the Pearson Correlation Coefficient without the use of deviations," Psychological Bulletin, January, 1917. This formula is sometimes referred to as Ayre's formula. He published it several years later.

correlation table. Thus if we should plot a correlation table for the relation between stature of children and their marks in an examination in arithmetic, we should find a low positive relation. We should find a relation between these two variables but it would not be close. The correlation would be positive but low. The reason for the positive correlation would be the fact that we had omitted to consider the age, which is related to stature. If we should plot a correlation table between the same two variables, namely, stature and the marks in the arithmetic examination for a group of children who are all of the same age, then the coefficient would probably be zero. The reason for this shift in the correlation coefficient is that in one case we keep a third variable, the chronological age, constant, whereas in the other case we allow it to enter the experiment uncontrolled.

Another factor that influences the size of the correlation coefficient is the linearity of the regression. If the line or diagonal tendency on the diagram is straight, or if it can be considered to be straight for practical purposes, the coefficient will be higher than if the diagonal tendency of the points on the scatter diagram fall in a curved line. For scatter diagrams of the latter sort the so-called Eta coefficient [1] should be determined, although it has the disadvantage that it is markedly affected by the size of the class interval chosen.

[1] For a discussion of the Eta coefficient or correlation ratio, see G. U. Yule, "An Introduction to the Theory of Statistics," Chapter X.

Still another factor that causes a low correlation coefficient is the restriction in the range of the variables. Suppose that we are plotting a correlation table between stature and age of children, and that we include in the table ages from zero to twenty. The relation will then be quite prominent. It will be clear at a glance that the older children are taller on the average than very young children. But now suppose that we restrict the range of the age on the correlation table to the limits 10 to 12 years. We shall then have no children recorded on the correlation table who are younger than 10 nor older than 12. The relation between stature and age, when age is restricted to a short piece of its total range, will be very low. If we divide the range of two years into twenty-four class intervals of one month each, and if we plot the corresponding statures to any suitable scale, we shall have a correlation table showing a low degree of relationship. One should always be on the lookout for this factor in judging correlations and correlation tables. For example, it is a matter of universal observation that the correlations between intelligence tests for college students and their scholastic standing are lower than the coefficients found in the high schools and the grade schools. The primary reason for this is probably that the college students represent, on the average, a selected group with a restricted range of mentality toward the upper end of the scale. If we should give an intelligence test to ten thousand people of the general population, as in the draft army, and then determine

what they could do in college, we should find a strikingly high coefficient of correlation because the lower end of the intelligence test scale would all be failures, including the total illiterates.

The amount of confidence that we give to coefficients of correlation should also be determined by the number of cases represented in the table. If the number of cases is small, such as twenty or thirty, and if the coefficient is relatively low, we should not place much confidence in the exact determination of the degree of relationship. The fact of presence or absence of relation may be fairly safely judged from a small number of cases, but the exact degree of relation should not be inferred unless the number of cases in the diagram is several hundred or several thousand. The degree of confidence to be placed in a correlation coefficient can be stated numerically in the form of a probable error which may either be calculated or determined from statistical tables.

The regression lines. Suppose that we know a man's stature to be 70 inches and that we want to guess his weight on the basis of a correlation table such as *Figure 38*. The best possible guess that we can give is the average weight of all the men who are 70 inches tall. We should then turn to the correlation table and determine the average weight of the men represented in the column for a stature of 70. There are 87 cases in that column and their average weight is about 150 pounds. We should therefore give 150 as the most intelligent guess of the weight of men who are 70 inches tall. If we should do this for

every vertical column on the diagram, we could plot a line connecting the averages of the columns. This line, or its equivalent best fitting straight line, is known as the *regression of weight on stature*, or the regression of y on x. Such a chart or line would enable us to make a prediction or an intelligent guess as to the weight of men when their stature is known, and our guess would be based solely on the known rough relation between these two variables.

Now suppose that the problem were reversed. We have discovered that men who are 70 inches tall weigh on the average 150 pounds. This may be taken as the normal or average weight for men of given stature. But now suppose that we want to guess the stature of a man who is known to weigh 150 pounds. We are not justified in saying that his stature would probably be 70 inches. The procedure would then be to determine the average stature of all the men who weigh 150 pounds. This is done by consulting the horizontal row in the correlation table representing the distribution of stature for men who weigh 150 pounds. There are 49 men represented in that class and their average stature is about 69 inches. We find therefore that the average weight for men who are 70 inches tall is about 150 pounds, but it does not follow that the average stature for men who weigh 150 pounds is 70 inches. In one case we are determining the average of a *vertical column* of entries in the correlation table, and in the other case we are determining the average of a *horizontal row* of entries in the table.

A line or curve may be drawn to represent the average of the rows and such a line would be called the *regression of height on weight*, or the regression of x on y. It is very essential to keep in mind which of the two regressions one is using in preparing statistical tables, and not to assume that the tables can be read in both directions. If x is known, a certain value should be given as the normal or average y, but if that same value of y is known, it would not be associated with the same value of x.

The regression equations are as follows:

Regression y on x (predicting y when x is known)

$$y = r\frac{\sigma_y}{\sigma_x}x$$

Regression x on y (predicting x when y is known)

$$x = r\frac{\sigma_x}{\sigma_y}y$$

In both of these equations, x and y denote deviations from the respective means.

Ordinarily we make predictions on the basis of the actual X and Y values, such as stature and weight, rather than on the basis of deviations of these measures from their means. The following forms will therefore be found useful in practical work:

Regression Y on X (predicting Y when X is known)

$$Y = r\frac{\sigma_y}{\sigma_x}X - \left[r\frac{\sigma_y}{\sigma_x}m_x - m_y\right]$$

Regression X on Y (predicting X when Y is known)

$$X = r\frac{\sigma_x}{\sigma_y}Y - \left[r\frac{\sigma_x}{\sigma_y}m_y - m_x\right]$$

In both of these equations, X and Y denote the actual original values of the two variables, and not the deviations from their respective means.

Problem 1. Plot a correlation table for any two variables and determine its correlation coefficient. Plot another correlation table representing only one-fifth of the total range of one of the variables. Calculate the coefficient for this new correlation table and show that its coefficient is smaller and why.

Problem 2. The following problem will probably afford opportunity for discussion of the regression lines. Assume that the average mental age of twelve-year-old children is twelve. Show why it is that the average age of children who have a mental age of twelve is not twelve.

Problem 3. Make a table of the proper weight for each stature as shown on any penny-in-the-slot scales. Determine your weight. Do these facts represent the regression of height on weight, or the regression of weight on height? Make some erroneous statements that could be made by a person who draws inferences from your table of height and weight without understanding what is meant by the regression lines.

Chapter Twenty-four

The Calculation of the Pearson Coefficient[1]

If one does not calculate the correlation coefficient often, one may find it necessary to return to the textbooks in order to relearn the method of computation. That should not be necessary if a data sheet is available with direct instructions for the computation separated from the discussion of the logic of the correlation table. I find that even those who calculate these coefficients frequently spread their calculations over many slips of paper and compute the various parts in more or less random order. It is almost impossible to check calculations that are not systematically arranged.

The correlation data sheet which I have been using is of double letter size so that it can be folded once and filed with letter-size papers and reports. I have arranged the sheet so that it should not be necessary to do any scribbling on loose slips of paper. A small space is provided for the scribbling that may be necessary. All items in the calculation, such as frequencies at the intersections of the arrays, sums, squares, and other numbers, belong in definite spaces on the data sheet. If one has occasion to return to the calculations, he can readily find what he is

[1] Reprinted from Journal of Educational Research, June, 1922, with the permission of the Editors.

looking for. The correlation data sheet is a labor-saving device.

Instructions for the use of the data sheet. In order to make these instructions brief and easily read I have indicated several of the columns by the numbers at the top of the sheet. (See *Figure 38.*) These numbers do not occur on the regular data sheet. They are inserted here for the purpose of these instructions only. The horizontal rows are labeled by letters for the same purpose. The rows and columns are known collectively as arrays. When I refer in these instructions to a certain square on the data sheet, I shall refer to it by the number of the column and the letter of the row which intersect in a square. Thus $C-3$ refers to a square on the sample data sheet which contains the figure 7. The sample set of data is drawn from West's textbook in statistics.[1]

1. Select any suitable class interval for the x-variable. On the data sheet these intervals are unity.

2. Record these class intervals in row A as shown (61, 62, etc.). Always arrange the small numbers to the left and the large numbers to the right.

3. Row B contains the upper class limits for the class intervals, which in the case of the example are 61.9, 62.9, etc. Row A contains, then, the lower class limit, and row B the upper class limit.

4. Select a suitable class interval for the y-variable. In the example this is five. Record the lower

[1] Carl J. West, "Introduction to Mathematical Statistics." Columbus, Ohio: R. G. Adams and Company, 1918. Page 67.

and upper class limits for these class intervals in columns 1 and 2 respectively. By tabulating both the upper and the lower class limits for each class interval one is sure that the actual range for each class interval is indicated on the data sheet. This serves to prevent confusion as to the square in which any particular frequency is to be recorded. Arrange the small numbers at the bottom and the large numbers at the top, as shown.

5. Record the frequencies in the squares of the correlation table. In the example these range between rows D and K and between columns 3 and 7.

6. Record in row L the sums of the frequencies in the vertical columns of the correlation table.

7. Record in column 8 the sums of the frequencies in the rows of the correlation table.

8. Add the numbers in the row L and record at N ($N =$) in the table of sums in the upper right-hand corner of the data sheet. In the example this sum is 750. This is the total number of cases.

9. Add the numbers in column 8 and see if the sum agrees with the N just recorded in the table of sums. It should be the same number.

10. Select any class interval as an arbitrary origin for the x-variable. In the example this is chosen at column 5. This arbitrary origin may be taken at any class interval but it reduces the arithmetical labor to choose it as close to the mean as possible. This should be done by inspection because it does not pay to calculate the mean specially for this purpose. Draw a blue pencil line or a wavy pencil line through

column 5 so as to mark off clearly at all parts of the correlation table the position of the arbitrary origin.

11. Do likewise for the y-variable. The arbitrary origin for the y-variable has been chosen at row G. This row is marked similarly to the corresponding class interval in x (column 5). The correlation table is now divided by the blue pencil lines into four quadrants. It should be remembered that these quadrants are marked off from the two assumed means and are not to be confused with the quadrants which would be marked off from the true means.

12. Record a zero at $M-5$. Record the successive class intervals on each side of this zero, as shown in the rest of row M. Do likewise in row C.

13. Record in row N the products fx. These are merely the products of the two rows immediately above N.

14. Record in row P the products fx^2. These are the products of the two rows immediately above P. This is easily verified by inspecting the sample data sheet.

15. Record a zero at $G-9$. Record above and below this the ascending and descending orders of class intervals, as shown in the rest of column 9.

16. Record the products fy in column 10. These are the products of the two columns immediately to the left of column 10.

17. Record the products fy^2 in column 11. The details of these calculations should be evident from the data sheet.

18. In row *D* you will find the numbers 3, 2, 2, 1 to the right of the assumed origin. Immediately above these numbers in row *C* you will find the numbers 1, 2, 3, 4. Each of the numbers in row *D* is to be multiplied by the number immediately above it in row *C*. These numbers are to be summed up thus:

$$(3 \times 1) + (2 \times 2) + (2 \times 3) + (1 \times 4) = 17$$

The number 17 is recorded at $D-13$. The sums that are obtained on the right-hand side of the assumed origin (column 5) are recorded in column 13. The sums similarly obtained on the left-hand side of the origin are recorded in column 12. This verbal statement will be clearer by noting visually on the data sheet the manner of the calculation.

This operation is repeated for each row with a separate sum for the right-hand side and for the left-hand side. For example: In row *F* we have the following sums:

Left side: $(3 \times 1) + (2 \times 2) + (2 \times 3) = 13$

Right side: $(10 \times 1) + (12 \times 2) + (11 \times 3) + (1 \times 5) + (1 \times 6) = 78$

These sums, 13 and 78, are recorded in the squares $F-12$ and $F-13$ respectively.

This operation can be performed with less effort on a comptometer or any other key-driven calculating machine. It is done as follows:

Consider the left side of row *F*. Set the finger on the 1-key at the lower right end of the machine. This key represents the deviation 1 of the *C*-row. Punch the key three times to represent the frequency

of three. Move the finger to key 2 immediately above your present location on the machine. Punch it twice to represent the frequency of two. Move the finger up one step again and punch the new key twice to represent the frequency of two. The sum of 13 is now on the machine without any further mental effort. This work can be done without looking at the machine, so that one may keep one's eyes on the correlation table.

Now consider the right side of row *F*. Place the finger on the 1-key at the lower right end of the machine. This key may be punched ten times or the key to its left may be punched once to represent the frequency of ten. Move the finger up one step on the machine. Punch this key twelve times or punch it twice and the key to its left once to represent the frequency of twelve. Move the finger up one step on the machine and punch the key eleven times or punch it once and the key to its left once. Move the finger up two steps on the machine and punch once; move the finger up once more and punch once. The sum of 78 is now in the machine and may be recorded at *F*−13. One will readily learn to move the finger one step at a time on the machine without looking at it even though he may not know more about the calculating machine. In this way the task is rendered easy even for a novice.

19. The difference between the two numbers 13 and 78 which were recorded at *F*−12 and at *F*−13 is 65. This number 65 is recorded at *F*−15. The same operation is carried out for all the figures in

columns 12 and 13. Their differences are recorded in columns 14 and 15. If the number in the 12-column is larger than the adjacent number in the 13-column, one records the difference in the 14-column. Similarly, if the number in the 13-column is larger than the adjacent number in the 12-column, the difference is recorded in column 15. One should visualize this right-and-left relation in order to remember it.

20. For the portion of the table above the assumed origin of the y-variable (row G) record the products of the corresponding numbers in columns 9 and 14 in column 16. Record the products of the corresponding numbers in columns 9 and 15 in column 17. These instructions apply above the assumed origin. Below the assumed origin the relation is reversed so that the products 9 and 14 are recorded in column 17 while the products 9 and 15 are recorded in 16. The following summary will make matters clearer:

	Columns
Above the assumed origin:	(9) × (14) = (16)
	(9) × (15) = (17)
Below the assumed origin:	(9) × (15) = (16)
	(9) × (14) = (17)

Another way to recall this is to note that columns 14 and 15 preserve the same spatial arrangement above the assumed origin but that below the origin this lateral arrangement is reversed.

Still another and perhaps the best way to recall this is to note that the y-deviations in column 9 are positive above the assumed origin and negative be-

low the assumed origin. This comes about from the fact that we plot the y-variable in column 1 with the high numbers at the top of the column and the small numbers at the bottom. If now the positive y-deviations in column 9 are multiplied by positive Σx in column 15, we obtain a positive Σxy and record the product in column 17. If we multiply the negative y-deviations of column 9 by the positive Σx in column 15, we obtain a negative Σxy and record the product in column 16. If we multiply the positive y-deviations of column 9 by the negative Σx in column 14, we obtain a negative Σxy, which we record in column 16. If we multiply the negative y-deviations of column 9 by the negative Σx in column 14, the product Σxy is positive and is recorded in column 17.

21. We are now ready to extract the sums for the final computation of the coefficient. Add the numbers to the left of the assumed origin in row N. Record the sum at $\Sigma x_{neg} =$ in the table of sums in the upper right-hand corner. It is 744.

22. Add the numbers to the right of the assumed origin in row N. Record the sum at $\Sigma x_{pos} =$ in the table. It is 669.

23. Determine the difference between Σx_{pos} and Σx_{neg} and record this difference at $\Sigma x =$ in the table of sums. It is -75. The relation is $\Sigma x = \Sigma x_{pos} - \Sigma x_{neg}$. Record the proper sign for the Σx.

24. Add the numbers in the row P on both sides of the assumed origin. Record this (4201) at $\Sigma x^2 =$ in the table.

25. Add the numbers above the assumed origin in column 10 and record the sum (880) at Σy_{pos} = in the table of sums.

26. Add the numbers below the assumed origin in column 10 and record the sum (933) at Σy_{neg} = in the table of sums.

27. Determine the difference between Σy_{pos} and Σy_{neg} and record the difference (− 53) at Σy = in the table of sums. The relation is $\Sigma y = \Sigma y_{pos} - \Sigma y_{neg}$. Record the proper sign for the Σy.

28. Add all the numbers in column 11. Record this sum (7377) at Σy^2 = in the table of sums.

29. Add all the numbers in column 16 and record the sum (zero) at Σxy_{neg} = in the table of sums.

30. Add all the numbers in column 17 and record the sum (2783) at Σxy_{pos} = in the table of sums.

31. Determine the difference between the Σxy_{pos} and the Σxy_{neg} and record the difference (+ 2783) at Σxy = in the table of sums. The relation is $\Sigma xy = \Sigma xy_{pos} - \Sigma xy_{neg}$. Record the proper sign for the Σxy.

The numbers that have been recorded in the table of sums may be used for substitution directly into the formulæ that are printed on the data sheet. From this point on one proceeds as shown by the formulæ until the coefficient is obtained. In the example the correlation coefficient is + .50.

The work of tabulating the correlation table and computing the coefficient for a problem like the one in the accompanying example requires about forty to sixty minutes after one becomes accustomed to the

sheet. If one is working with the correlation coefficients continually, the calculation can be made in less time. Verbal description of spatial relations is necessarily clumsy. To read these instructions without consulting the data sheet might give one the impression that these calculations are awkward and even intricate; but after the necessarily involved verbal description has been translated into visual or graphic terms, the work can be done by a clerk who knows nothing about correlation statistics.

Chapter Twenty-five

Correlation by Ranks

When two variables are expressed in terms of their *ranks* and not in terms of their *original values* or *scores*, one can find only approximately the coefficient of correlation between the two variables. The product moment, or Pearson, coefficient cannot be applied to a set of paired ranks as it is applied to a set of paired measurements. When the only data available are in the form of ranks, the correlation coefficient for the two variables can be found approximately by formulæ that differ from the product moment formulæ previously discussed. Sometimes the rank formulæ are preferred even when the available data contain the original values or measurements, but such preference is based on the relative ease of calculation of the rank coefficient as compared with the product moment coefficient.

The correlation coefficient for paired ranks is given by the following relation:

$$\rho = 1 - \frac{6\,\Sigma(k_x - k_y)^2}{n(n^2 - 1)} = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)},$$

in which ρ = the rank correlation coefficient
r = the product moment correlation coefficient or its equivalent

n = the number of cases
k_x = the ranks of the x-variable
k_y = the ranks of the y-variable
d = the difference between ranks of paired measurements

When a set of paired ranks is used for determining the correlation coefficient for the two variables, one is making the assumption that the distributions of the two variables are normal; in many instances this is a safe assumption for practical purposes.

In *Table 17* we have an example of the application of the rank correlation method. In the first column there are listed the twenty-four individuals in the group. The constant n is therefore 24. In the second and third columns we have the original measurements in the two variables for each of the twenty-four individuals. These variables represent test scores. In the fourth column have been listed the ranks of the individuals in *Test X*. For example, the individual 20 has a score of 17 in Test X, and since that is the lowest score in the whole list, he is given an absolute rank of 1, which is recorded in the fourth column. Similarly, individual 17 has a score of 198 in the same test, and since that is the highest score in the group, he is given the highest rank, namely, 24, as shown in the fourth column. In fact, the fourth column contains the absolute rank of the numbers entered in the second column. When there are two or more identical scores, it is necessary to balance the ranks, as described in previous chapters

on ranks. In the fifth column we have the absolute ranks of the numbers entered in the third column.

The next step is to list in the *d* column the differences between the ranks as recorded in the fourth and fifth columns. For example, the difference between ranks 12 and 14 for the first individual is 2 and that is recorded in the *d* column irrespective of sign. In the last column the squares of the *d* column are entered. Since it is the squares of the differences in rank that enter into the correlation formula, it is unnecessary to take any account of the signs of the rank differences.

When the tabulation has been made, the last column is summed. The remaining operations consist merely in substituting the numerical values in the formula, as shown in *Table 17*. The rank correlation coefficient is usually designated ρ; in this example its value is + .58.

For a distribution which is approximately normal one can translate the value of ρ into an approximate value for r by the following relation:

$$r = 2 \sin \frac{\pi}{6} \rho$$

In order to facilitate this transposition from ρ to r *Table 18* has been prepared. By this table one can obtain an approximate estimate of the product moment coefficient of correlation from the rank coefficient as calculated from the absolute ranks.

Problem 1. Calculate the product moment coefficient of correlation for the data in *Table 17* and compare its value with

The Individual Members of the Group	*Their Scores in X*	*Their Scores in Y*	*Their Ranks in X*	*Their Ranks in Y*	*Difference Between Ranks*	*Difference Squared*
#	X	Y	k_x	k_y	d	d^2
1	104	18	12	14	2	4.00
2	83	14	9	8	1	1.00
3	155	17	20	12.5	7.5	56.25
4	165	23	21	20.5	.5	.25
5	178	19	22	15.5	6.5	42.25
6	147	25	18	23	5	25.00
7	22	12	2	5	3	9.00
8	84	10	10	1.5	8.5	72.25
9	27	19	3	15.5	12.5	156.25
10	63	21	7	18	11	121.00
11	117	17	14	12.5	1.5	2.25
12	94	23	11	20.5	9.5	90.25
13	74	16	8	11	3	9.00
14	59	13	5.5	7	1.5	2.25
15	118	15	15	9.5	5.5	30.25
16	131	21	16	18	2	4.00
17	198	24	24	22	2	4.00
18	108	12	13	5	8	64.00
19	59	11	5.5	3	2.5	6.25
20	17	15	1	9.5	8.5	72.25
21	29	10	4	1.5	2.5	6.25
22	189	21	23	18	5	25.00
23	152	26	19	24	5	25.00
24	146	12	17	5	12	144.00

$$\Sigma(k_x - k_y)^2 = 972$$

$$6\,\Sigma(k_x - k_y)^2 = 5832$$

$$n = 24$$

$$n(n^2 - 1) = 13{,}800$$

$$\rho = 1 - \frac{6\,\Sigma(k_x - k_y)^2}{n(n^2 - 1)} = +.58$$

Transposing from Table 18 we have:

Equivalent value for $r = +.60$

Table 17. Calculation of correlation coefficient by ranks

the value for ρ and the estimated value of r as determined by *Table 18.*

ρ	r
.00	.000
.05	.052
.10	.105
.15	.157
.20	.210
.25	.261
.30	.313
.35	.364
.40	.416
.45	.467
.50	.518
.55	.568
.60	.618
.65	.668
.70	.717
.75	.765
.80	.813
.85	.861
.90	.908
.95	.954
1.00	1.000

Table 18. Values of Pearson coefficient of correlation corresponding to various values of the rank correlation coefficient

APPENDIX

Appendix

Ordinates of the Probability Curve

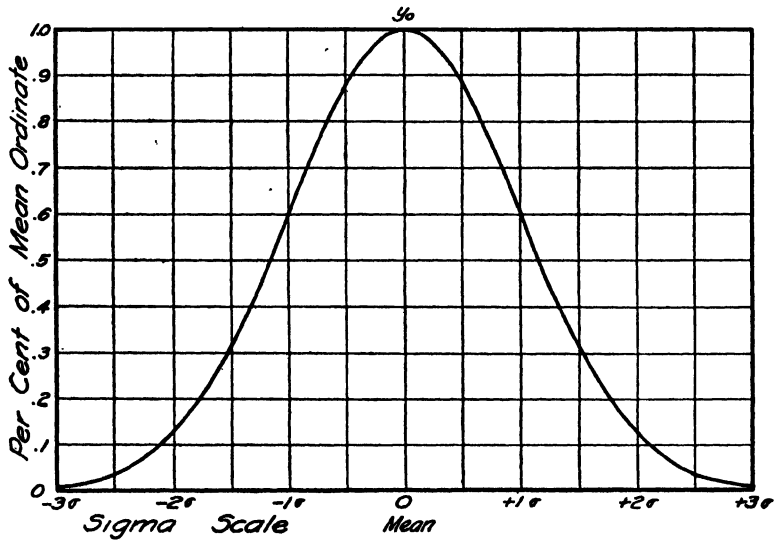


Figure 39

The probability curve. The mean is designated zero on the sigma scale. The highest ordinate, y_0, is at the mean. This ordinate can be computed for any particular distribution by the following relation:

$$y_0 = \frac{n}{\sigma\sqrt{2\pi}} = \frac{n}{2.5066\sigma}$$

All the other ordinates are expressed as fractions of the mean ordinate depending on their distance from the mean ordinate.

Example. The ordinate which is + 1 σ from the mean, either above or below, is .606 of the mean ordinate. The ordinate which is 1.34 σ from the mean is .407 of the mean ordinate. Verify these facts in the opposite table and also in the above figure.

It is necessary to translate the x-scale into corresponding σ values before a probability curve can be plotted. Thus, if a score is 65, σ = 10, m = 50, the score 65 is said to be + 1.5 σ on the sigma scale. Its ordinate on the probability curve would be .325 of y_0.

Sigma	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	1.00000	.99995	.99980	.99955	.99920	.99875	.99820	.99755	.99680	.99596
0.1	.99501	.99397	.99283	.99159	.99025	.98881	.98728	.98565	.98393	.98211
0.2	.98020	.97819	.97609	.97390	.97161	.96923	.96676	.96421	.96156	.95882
0.3	.95600	.95309	.95009	.94701	.94384	.94059	.93725	.93384	.93034	.92677
0.4	.92312	.91939	.91558	.91169	.90774	.90371	.89960	.89543	.89119	.88688
0.5	.88250	.87805	.87354	.86897	.86433	.85963	.85487	.85006	.84518	.84025
0.6	.83527	.83023	.82514	.82000	.81481	.80957	.80429	.79896	.79358	.78816
0.7	.78270	.77721	.77167	.76609	.76048	.75484	.74916	.74345	.73771	.73194
0.8	.72615	.72033	.71448	.70861	.70272	.69680	.69087	.68492	.67896	.67297
0.9	.66698	.66097	.65495	.64892	.64288	.63683	.63078	.62472	.61866	.61259
1.0	.60653	.60047	.59440	.58834	.58228	.57623	.57018	.56414	.55811	.55209
1.1	.54607	.54007	.53409	.52811	.52215	.51620	.51028	.50437	.49848	.49260
1.2	.48675	.48092	.47511	.46933	.46357	.45783	.45212	.44644	.44078	.43516
1.3	.42956	.42399	.41845	.41294	.40747	.40202	.39661	.39123	.38589	.38058
1.4	.37531	.37007	.36488	.35971	.35459	.34950	.34445	.33944	.33447	.32954
1.5	.32465	.31980	.31500	.31023	.30550	.30082	.29618	.29158	.28702	.28251
1.6	.27804	.27361	.26923	.26489	.26059	.25634	.25213	.24797	.24385	.23978
1.7	.23575	.23176	.22782	.22392	.22007	.21627	.21250	.20879	.20511	.20148
1.8	.19790	.19436	.19086	.18741	.18400	.18064	.17732	.17404	.17081	.16762
1.9	.16447	.16137	.15831	.15529	.15232	.14938	.14649	.14364	.14083	.13806
2.0	.13534	.13265	.13000	.12740	.12483	.12230	.11982	.11737	.11496	.11259
2.1	.11025	.10795	.10570	.10347	.10129	.09914	.09702	.09495	.09290	.09090
2.2	.08892	.08698	.08508	.08320	.08137	.07956	.07779	.07604	.07433	.07265
2.3	.07101	.06939	.06780	.06624	.06471	.06321	.06174	.06030	.05888	.05750
2.4	.05613	.05480	.05349	.05221	.05096	.04972	.04852	.04734	.04618	.04505
2.5	.04394	.04285	.04179	.04074	.03972	.03873	.03775	.03679	.03576	.03494
2.6	.03405	.03317	.03232	.03148	.03066	.02986	.02908	.02831	.02757	.02683
2.7	.02612	.02542	.02474	.02408	.02343	.02279	.02217	.02157	.02098	.02040
2.8	.01984	.01929	.01876	.01823	.01772	.01723	.01674	.01627	.01581	.01536
2.9	.01492	.01449	.01408	.01367	.01328	.01289	.01252	.01215	.01179	.01145
3.0	.01111									
3.1	.00819									
3.2	.00598									
3.3	.00432									
3.4	.00309									
3.5	.00219									
3.6	.00153									
3.7	.00106									
3.8	.00073									
3.9	.00050									
4.0	.00034									
5.0	.000004									

Table 19. Ordinates of the Probability Curve[1]

[1] Adapted from tables calculated by W. F. Sheppard and published in "Tables for Statisticians and Biometricians," edited by Karl Pearson. Cambridge University Press.

Areas of the Probability Surface

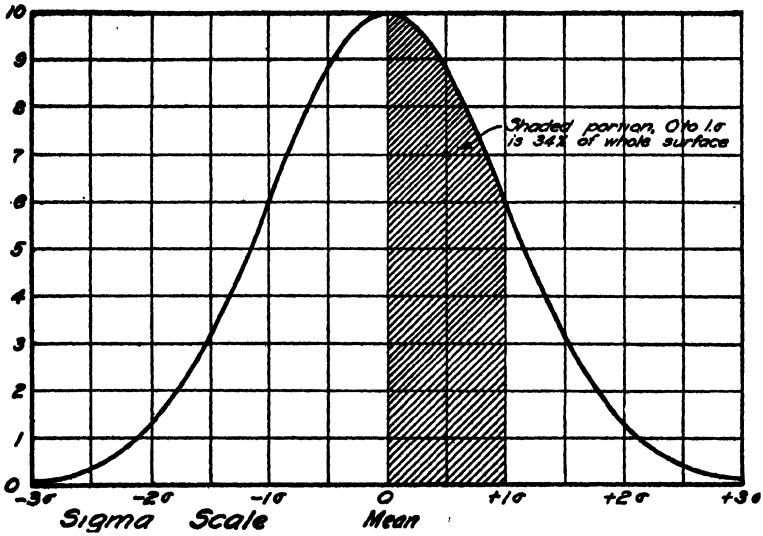


Figure 40

The mean is designated zero on the sigma scale. All other points on the scale are designated by their sigma value in *Figure 40*. The table shows the proportion of the whole surface that lies between the mean and any specified sigma point on the scale.

Examples. The proportion of the whole surface between the mean and $+1\sigma$ is .3413. The proportion of the whole surface between the mean and -1.65σ is .4505 or 45%. The proportion of the whole surface between $+1\sigma$ and -65σ is the sum of these two percentages, or 79%.

Assume that a probability surface has a mean of 50, a standard deviation of 10, and that we want to know the proportion of the whole surface between the mean 50 and the point 65. The point 65 is expressed as $+1.5\sigma$ and, according to *Table 20*, the proportion of the whole surface that lies between the mean and $+1.5\sigma$ is .4332, or approximately 43%.

Sigma	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.00000	.00399	.00798	.01197	.01595	.01994	.02392	.02790	.03188	.03586
0.1	.03983	.04380	.04776	.05172	.05567	.05962	.06356	.06749	.07142	.07535
0.2	.07926	.08317	.08706	.09095	.09483	.09871	.10257	.10642	.11026	.11409
0.3	.11791	.12172	.12552	.12930	.13307	.13683	.14058	.14431	.14803	.15173
0.4	.15542	.15910	.16276	.16640	.17003	.17364	.17724	.18082	.18439	.18793
0.5	.19146	.19497	.19847	.20194	.20540	.20884	.21226	.21566	.21904	.22240
0.6	.22575	.22907	.23237	.23565	.23891	.24215	.24537	.24857	.25175	.25490
0.7	.25804	.26115	.26424	.26730	.27035	.27337	.27637	.27935	.29230	.28524
0.8	.28814	.29103	.29389	.29673	.29955	.30234	.30511	.30785	.31057	.31327
0.9	.31594	.31859	.32121	.32381	.32639	.32894	.33147	.33398	.33646	.33891
1.0	.34134	.34375	.34614	.34850	.35083	.35314	.35543	.35769	.35993	.36214
1.1	.36433	.36650	.36864	.37076	.37286	.37493	.37698	.37900	.38100	.38298
1.2	.38493	.38686	.38877	.39065	.39251	.39435	.39617	.39796	.39973	.40147
1.3	.40320	.40490	.40658	.40824	.40988	.41149	.41309	.41466	.41621	.41774
1.4	.41924	.42073	.42220	.42364	.42507	.42647	.42786	.42922	.43056	.43198
1.5	.43319	.43448	.43574	.43699	.43822	.43943	.44062	.44179	.44295	.44408
1.6	.44520	.44630	.44738	.44845	.44950	.45053	.45154	.45254	.45352	.45449
1.7	.45543	.45637	.45728	.45818	.45907	.45994	.46080	.46164	.46246	.46327
1.8	.46407	.46485	.46562	.46638	.46712	.46784	.46856	.46926	.46995	.47062
1.9	.47128	.47193	.47257	.47320	.47381	.47441	.47500	.47558	.47615	.47670
2.0	.47725	.47778	.47831	.47882	.47932	.47982	.48030	.48077	.48124	.48169
2.1	.48214	.48257	.48300	.48341	.48382	.48422	.48461	.48500	.48537	.48574
2.2	.48610	.48645	.48679	.48713	.48745	.48778	.48809	.48840	.48870	.48899
2.3	.48928	.48956	.48983	.49010	.49036	.49061	.49086	.49111	.49134	.49158
2.4	.49180	.49202	.49224	.49245	.49266	.49286	.49305	.49324	.49343	.49361
2.5	.49379	.49396	.49413	.49430	.49446	.49461	.49477	.49492	.49506	.49520
2.6	.49534	.49547	.49560	.49573	.49585	.49598	.49609	.49621	.49632	.49643
2.7	.49653	.49664	.49674	.49683	.49693	.49702	.49711	.49720	.49728	.49736
2.8	.49744	.49752	.49760	.49767	.49774	.49781	.49788	.49795	.49801	.49807
2.9	.49813	.49819	.49825	.49831	.49836	.49841	.49846	.49851	.49856	.49860
3.0	.49865									
3.1	.49903									
3.2	.49931									
3.3	.49952									
3.4	.49966									
3.5	.49977									
4.0	.49997									
5.0	.4999997									

Table 20. Areas in the probability surface[1]

[1] Adapted from tables calculated by W. F. Sheppard and published in "Tables for Statisticians and Biometricians," edited by Karl Pearson. Cambridge University Press.

INDEX